The Development of Writing Competence in Grade Nine Papua New Guinea High School Students: an investigation of the relationship between personal history narrative, imagined story narrative and persuasive writing

Angie Phillip

PhD Thesis

University of Edinburgh

1998

DECLARATION

I declare that this thesis was composed by myself and that the work therein is my own.

Angie Phillip
September 1999

ABSTRACT

The development of writing competence in Papua New Guinea Grade Nine high school students was described in order to investigate the transition from narrative to argumentative or persuasive writing. The study used a pretest/posttest method and scripts were scored holistically and described according to objective measures (t-unit measures, fluency scored by number of words per timed essay, and accuracy described by measures of error per 100 words). Narrative writing was hypothesised to fall into three categories since it seemed that different cognitive processes were required for their production, and practice in two of these formed the treatment. A control group was given practice in personal history narrative, while an experimental group was given practice in imagined story narrative. The first objective was to investigate the relationship between the three types of writing, and the hypothesised hierarchy of difficulty, where persuasive writing was more difficult than imagined story narrative, which was, in turn, more difficult than personal history narrative, was confirmed. The second objective was to chart the development of writing competence over three quarters of an academic year. The writing of almost all the students improved to some extent and the improvement was marked by a significant increase in fluency in all three writing types. Patterns of error, however, varied between the types of writing. As competence increased in both types of narrative writing, overall error decreased, while improvement in persuasive writing appeared to be associated with a slight increase in error. In all three types of writing the proportion of spelling errors increased as competence developed, while the proportion of errors to do with coherence and cohesion fell. The third objective was to investigate the effect on the development of writing competence of practice in imagined story narrative, as opposed to the effect of practice in personal history narrative. Writing types had been mixed to some extent, both during the treatment and during the tests, so the experiment actually compared practice in more of a particular kind of narrative than exclusive practice in that type. The group who had received practice in imagined story narrative showed a significantly increased performance over the control group in persuasive writing, according to the holistic ratings. Holistic ratings, however, are difficult to rely on, as shown by a post hoc re-evaluation of scripts, but the better performance in persuasive writing of the experimental group was supported by a significantly greater increase in accuracy, measured objectively. It is argued that the students who received the imagined story narrative practice were able to write more accurate persuasive essays because practice in imagining had prepared them for some of the requirements of persuasive writing.

ACKNOWLEDGMENTS

I would like to thank my late father, Charles Herbert Gee, and my mother, Mabel Gee, for the constant encouragement they have given me to complete this thesis. Without their support and the help given by my late aunt, Susan Gee, and by the staunch friend of our family, Ronald Adey, this study would not have been possible. I would like, too, to thank my son, Jay Albus, who has offered support and good humour throughout. At Edinburgh University I am indebted to both my supervisors, Keith Mitchell and Tony Lynch, who have not only provided excellent advice, but who have also been consistently encouraging and sympathetic. Special thanks is due to Kate Marshall for her kindness, her help and her immediate response to queries during stressful times. In addition I would like to thank Jacqueline Gollin and Cathy Benson for their helpful advice on many occasions. In Papua New Guinea, I would like to thank the raters who marked hundreds of essays with no payment: Sam Sabiyam, Helen Rutstein, Barbara Burrows, Nuegu Billy, Moresby Goasa, Ephraim Moguna, Canitz Phillip, Teresa Tokanini, Menser Ragin, Julie Setu and Janet Armin. I would like, too, to thank the staff of Laloki High School for their support during the project. During the last few months I have been indebted to Bonnie Tauwala at the University of Papua New Guinea (UPNG) for coordinating the updating of information for the thesis. She has achieved this despite power cuts, student riots and a list of other difficulties too extensive to include here. For updating of information I am indebted to Ruth Ray, Peter Woods and David Kelly from the PNG Department of Education and to Nicolas Faraclas from the Department of Language and Literature at UPNG. At Oxford Brookes University, I would like to thank the School of Languages for giving me a term's sabbatical in 1997 to help me complete the study and I would like to thank Professor Sean Hand, Head of School, Mary Anne Ansell, my Head of Department and my greatly appreciated colleagues: Richard Haill, Colin Connolly, Clare McKinley, Liz Sayigh, Teresa Woodbridge, Lynn Errey, Ruth Green, Nicki Parsons and Annie McCartan for the unstinting support they have given me despite their own heavy workloads. I would like to thank Canitz Phillip for early help with computer matters and Dan Robertson for his help with statistics. I would like, too, to thank Itsuki and Maki Matayoshi for keeping me constantly supplied with music, food and flowers and Peter Galpin for commenting on numerous early drafts. Most recently and most of all, I am indebted to Paul Way-Rider for his invaluable help with tables, with data, and for final editing for without his help I would not have been able to complete the task. Finally, I would like to thank the grade nine students of Laloki High School in 1990, who were not only the subjects of the study, which implies some sort of passive role, but who contributed actively and substantially to any insights it may possess.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

eft - error-free t-unit

ESL - English as a Second Language

EFL - English as a Foreign Language

ISN - imagined story narrative

L1 - first language (mother tongue)

L2 - second or subsequent language

NS - native-speaker of English

NNS - non-native speaker of English

OPN - other people's narrative

PHN - personal history narrative

PNG - Papua New Guinea

PW - persuasive writing

UPNG - University of Papua New Guinea

STM - short term memory

TESL - Teaching English as a Second Language

TWE - Test of Written English

PART 1: BACKGROUND AND RESEARCH DESIGN

CHAPTER 1 - INTRODUCTION

At the beginning of the 1990's and probably well before that time, most of the new entrants to the University of Papua New Guinea, drawn from the cream of the country's high school elite, were arriving at university unable to write academic prose. Somehow the problem had to be solved. The students had to be taught how to write cogent argument, to anticipate and deal with counter-argument and to organise their writing so that it made sense. In order to help solve the problem, I decided to study the writing of high school students to see how writing competence developed. The intention was to try and map the transition from narrative writing, an easy genre, to persuasive writing, a difficult one.

This chapter will describe the research setting and the recent history of writing pedagogy in Papua New Guinea (PNG). The aims of the research will be explained as well as the reasons for choosing to study the relationship between the types of writing that were considered most appropriate for the investigation of the development of writing competence.

1.1 Research setting - Papua New Guinea

This section is intended to put the research questions in context. It is hoped that any insights from the research may be relevant to other situations where English is learned as a second language and maybe also to first language situations, but any conclusions need to be rooted in an awareness of the particular situation in which the research was carried out. In particular the constraints on the research experiment such as very limited teaching and study time for the students, the lack of materials, the workload of PNG students, who are required to do heavy physical work each day in addition to their studies, need to be taken into account.

1.1.1 Location

Papua New Guinea is a tropical country. Its nearest neighbours are Australia to the south and Indonesia to the west. There is a close but uneasy relationship with Australia. The unease is caused partly by fairly recent memories of Australians as rulers (until independence in 1975) and partly by PNG's present dependence on Australian aid money. The cultural influence of Australia, however, is

significant, undeniable, and with the advent of television, increasing. Indonesia, on the other hand, has remained until recently a rather distant neighbour although it occupies the other half of the large island land mass where PNG is situated. In recent times there has been an attempt to forge a closer relationship with Indonesia as a result of the growing awareness on the part of both PNG and Australia of the importance of south east Asian countries as trading partners. PNG is mostly mountainous with dense rainforest, which has been preserved so far only because of the difficult terrain which makes logging hard and dangerous. There are few roads and there is no road-link at all between the north and the south of the country. Round the coast there are long white beaches with palm trees alternating with steamy mangrove swamps.

1.1.2 Language, literacy and education

Papua New Guinea is linguistically and culturally rich. There are 869 languages in use (Ahai & Faraclas 1993) and it is the norm for a person to speak three languages or more. Language learning is considered to be easy and there is a general air of surprise that anyone should consider it difficult. The three national languages of Papua New Guinea are Tok Pisin, Hiri Motu and English. English is usually the second, third or sometimes fourth language that a student learns, but it is the language that is used for education, for most communication in the public service, and for many television programmes. Television at the time of the study was available in only a few urban areas, although the broadcasting area is gradually being extended. Patterns of language use are rapidly changing and English is used less than previously for government debate, for television advertisements and news items, and not usually for the social intercourse of any group. National as well as local radio stations produce many programmes in languages other than English. Tok Pisin in particular is spreading rapidly, has become a creole, and is the dominant oral lingua franca. Hiri Motu, on the other hand, seems to be declining in importance.

While almost everyone seems to display high levels of oral proficiency in several languages, literacy is another matter. The proportion of adult literacy was quoted as 46% (Asian Development Bank 1991), but workers in the literacy field believe that the figure was inflated. Papua New Guinea is committed to a policy of free education for all, but the money to achieve this is not yet available. Primary school

enrolment was reported as being 64% in 1991, but secondary school enrolment was only 14% (Asian Development Bank 1991). In addition, there are few reading materials available in any language. At the time of the study there were only one or two national newspapers. The most popular national English language daily was *The Post Courier,* which was read every day by most educated people including high school students, whereas *The Times of Papua New Guinea,* a weekly broadsheet tended to be read by only a small proportion of professional people. In addition to these, there was a weekly Tok Pisin newspaper, but few other reading materials. It came as no surprise, therefore, when Sulaiman (1990) reported that 70% of the students at the PNG University of Technology said their preferred reading material was newspapers. Bookshops and libraries were, and still are, virtually non-existent in most parts of the country so that high school students tend to be totally dependent on their school libraries, which usually have very small collections of books and limited borrowing facilities. Primary schools are called 'community schools' in Papua New Guinea and community school students often have no opportunity at all to either buy books or to borrow them. There are very few books available in languages other than English, although literacy workers are making strenuous and successful efforts to help communities to become literate in their mother tongue and to produce books in the local language.

The language policy for education is decided upon and prescribed by the National Department of Education and until very recently instruction from preschool to university has been entirely through the English language. Current planning intends that English should remain the dominant language of instruction, but should, at primary and secondary levels, be based upon and be supported by, teaching in a language chosen by the local community. Children in PNG start community school at approximately seven years of age and most finish their education at the end of grade six, although some continue to grade ten and a few to grade twelve. Age ranges for particular grades are not as fixed as they usually are in the West, since a student might have time out of school between grades while the family seeks to raise another year's school fees. Table 1 below shows the structure of the school system.

Table 1: Structure of the school system in Papua New Guinea

Age (approx.)	Grades	Type of School	Medium of Instruction
6-8 yrs		Tokples School *	local language
7-12 yrs	1-6	Community School	local language & English
13-16 yrs	7-10	Provincial High School	mainly English
17-18 yrs	11-12	National High School	English

Key: * Tokples Schools are available only in some areas and are run by members of the community. The schools are designed to offer literacy and elementary schooling in the local language with the aim of giving the pupils a sound start to their education before they transfer to community school. (Reform is currently planned where the first three years of formal schooling will be in 'Elementary Schools' using vernacular languages.)

The situations where English is still the sole medium of instruction are expected to change as new language policies are introduced into the education system. The new policies endorse multiple language use in the classroom, especially at the base of the education system, but practical difficulties can be expected to slow the process. Even when the changes are fully implemented, the main medium of instruction is intended to be English. For the students in this study, English had been the medium of instruction throughout their school life.

1.1.3 Stories and sermons

A favourite pastime in Papua New Guinea is 'telling stories'. If you ask students what they were doing in their free time, the most usual answer is 'telling stories'. To 'tell stories' in this sense means to exchange accounts of personal experience, or to gossip. It does not mean to tell fictional stories. The other kinds of story that play a large part in the cultures of PNG are myths and legends. These kinds of stories are relatively fixed and are not thought of as 'stories' in the way stories would be understood in the West. They are passed down from generation to generation through an oral tradition, and they serve not only to entertain but to explain why things are the way they are: they teach a view of the world. Another more recent kind of story that cannot be changed and that is associated with 'truth' and therefore goodness, is the bible story and following on from it, the religious sermon exhorting people to good behaviour. On the other hand, the comparatively recent introduction into some parts of

the country of cinema and TV movies that tell fictional stories are commonly considered to be a bad influence, especially on young people.

In contemplating the speech styles and values that form the starting points for acts of literacy for PNG children, it is important to recognise the relationship between literacy and orality. This relationship is complex and variable as shown, for example, in the powerful poetic account by Benterrak, Muecke & Roe (1984) of 'reading' the Roebuck Plains country of N.W Australia, where 'reading' means 'understanding' and where the text communicates in a variety of ways that cross the usual academic conventions to include accounts of aboriginal speech and story-telling presented together with pictures. The relationship between speech and writing shows overlap rather than an absolute difference (Biber 1988). although it seems that linguistic relationships between registers demonstrate some stable similarities, even across languages as diverse as English, Nukulaelae, Tuvaluan, Korean and Somali (Biber 1995). More important, perhaps, than register variation between the PNG cultures and the target English medium culture is the fact that the *values* associated with oral practices of the home culture often transfer, at least to some extent, to written practices in the target culture. In Papua New Guinea the style, content and values commonly associated with oral practices are frequently evident in pieces of writing, despite the fact that the writing is usually not in the mother tongue but in English. Not only is an oral style often evident in written stories, but the frequently heckling style used in sermons and village speeches can be seen, too, both in school-generated persuasive essays, as well as in letters to the editors of newspapers. The values of traditional oral practices, such as the value attached to the 'truth' of stories, is also transferred to written practice as far as is possible. The 'truth' values attached to story telling make sense when one considers the traditional importance of the oral narratives of myths and legends which explain the world to PNG children. An extension of the explanations provided by myths and legends are those provided by bible stories and such stories have been treated with similar respect and care. An awareness of the historical intent of acts of literacy is important in understanding their role as social action (Tonkin 1992). She stresses the fact that 'memory makes us' (1992:117), that it forms both our personal and social identity.

Narrative has been seen as central to a child's development of communicative ability and, following from this, crucial to the way a society expresses itself and communicates both in speech and in writing (Halliday 1976, Scollon 1976, in Scollon & Scollon 1981). An understanding of differences where two sets of styles and values collide, as they did, for example, between native Alaskans and whites in Scollon and Scollon's (1981) study of discrimination in the courts of Alaska, and as they do in PNG where the village style and the school style are different, is obviously helpful in enabling communication. As early as 1978 Cazden and Hymes (cited in Scollon & Scollon 1981) pointed out that within American education, and this applies to British and Australian education, too, there is a widespread bias against narrative as a communicative medium. It follows then that children from societies that favour a narrative communicative style are disadvantaged when confronted with the expectations of mainstream English medium education. The conflict becomes even greater when it emerges that the expectation of narrative content as well as of the value attached to it differs, too, as it does in PNG, between the education system and the teachers and children involved in it. For example, it is not part of PNG culture to invent stories. Such a practice would be generally regarded as 'telling lies', yet the syllabus now requires this.

There are many societies like those in Papua New Guinea, that require writing to be 'truthful' and 'authoritative'. The Nukulaelae islanders, for example, consider that literature should be 'truthful'. 'Children are socialised early to accept the written word as the bearer of truth, to be memorised and recited in the appropriate context: once a year in particular, on Children's Sunday...' (Besnier 1995:164) The inhabitants of Roadville in Brice Heath's (1983) study set great store by the religious and literary 'truth' of what they encourage their children to read. A protective attitude towards knowledge, an emphasis on memorisation and repetition of texts and a literacy closely associated with religion are characteristics of what Goody describes as 'restricted literacy', a term which 'refers to situations in which literacy is "restricted by factors other than the techniques of writing itself"' (Goody 1977 in Besnier 1995:170). It is important to appreciate, however, that even when general terms such as 'restricted literacy' seem to apply, that literacy is 'a fundamentally heterogeneous phenomenon' (Besnier 1995:12) which is meaningful only in terms of its own socio-cultural context (Brice-Heath 1983; Besnier 1995). Social values focussed upon through acts of reading and writing

can and do vary from one society to another despite superficial similarities. For example, the Nukulaelae islanders in Besnier's study valued literacy primarily as a means of communicating 'God's word' and enabling people to read the bible, whereas PNG people would, on the whole, value literacy more as a means of education and ultimately self betterment in a material sense. The three groups of people in Brice Heath's (1983) southern USA study each valued and used literacy in different ways: the educated middle-class townspeople used literacy mainly to encourage their children to imagine and question, the white working class community of Roadville used literacy to teach their children acceptance of moral values, while the black working class community of Trackton used literacy as a functional tool for dealing with wider society while concentrating on encouraging imaginative acts of orality for child development (especially for the boys).

The differences in viewpoint and value associated with acts of literacy in various societies can and do cause conflict when the oral and written practices of the receivers of education differ markedly from those used by the education providers. Besnier makes the point that literacy is often not used by the receivers in the way it was intended by the introducers, and cites Gewertz and Errington's (1991 in Besnier 1995) study of the Chambri people of the Middle Sepik region of PNG. The Chambri were introduced to literacy in the context of government efforts to bring development to the region, but the people used it, too, for their own purposes, particularly to create documents to demonstrate status which they could use for personal advancement. A conflict arises in a more general sense in PNG in that both educators and parents want the children to do well in the education system so that they can be successful and earn lots of money, and yet the teachers are often required to transmit values that are alien, as noted above. A good example of this is the discussion that has gone on about whether creative i.e. imaginative writing should be taught in schools or not (where imaginative writing is valued by the western education system, but where factual repetition and respect for unchanging authoritative accounts, is valued traditionally).

In the past, occasional suggestions from expatriate educators were made along the lines that it would not be appropriate to introduce creative writing, in the sense of invented writing, into PNG schools because such a practice would run contrary to cultural traditions. The Papua New Guinea educators

seemed until recently to have come to the same conclusion but have given a different reason for their decision. Creative writing was excluded for many years from all but the national high schools, which ran grades eleven and twelve, because of the view that it would not be directly useful for students. Recently, since the data for this research was collected in 1990, the thinking has changed and the current view is that creative writing is useful and should be taught. The 'story in school', however, carries different expectations from 'the story in the village' and this, as it did for Brice Heath's (1983) Trackton and Roadville working-class students, can create difficulties. Students have to learn new definitions, new styles and structures, and, most difficult of all, new values. School, for a Papua New Guinea child, almost always involves emotional conflict. The difficulty in community school is that a new language with its different set of cultural values has to be learned and this strikes at the heart of a child's identity and self esteem. If a prized place in secondary school is obtained, it will usually involve the sacrifice of leaving home and going far from family and familiar faces.

1.1.4 Secondary school student life

The relatively small number of high schools scattered in the large geographical area that the country occupies together with the difficulty of travel because of poor or non-existent roads and no railway system means that secondary school education is conducted almost entirely in boarding schools. Going to secondary school can involve a two-week trek over the mountains in order to get there, or a long canoe journey. For many students an airstrip is required. Some students spend a year at a time away from their relatives because of the difficulty or the expense of travel. Students whose schools are near to their families are few and fortunate and receive weekend and term breaks, when they can leave the school to visit their families.

Going to school in PNG involves a sense of both privilege and sacrifice. Extended families sacrifice precious savings in order to pay school fees, and the students are extremely aware of this. To gain a place in secondary school, it is necessary not only to do well in the Grade Six Examination, but also to have family members who will scrape together large amounts of high school fee money. The students pay the price for their privilege by carrying the burden of responsibility to meet their parents' expectations. At the end of their education they will be expected to repay their debt to the extended

family. Students frequently pay an additional emotional price of separation from their families in order to learn.

1.2 Writing pedagogy in Papua New Guinea

1.2.1 National examinations

During the 1980's the Department of Education put an increasing emphasis on the development of writing skills. Despite initial objections from the measurement services section, who complained that essay writing was not capable of reliable assessment, a component called 'Written Expression' was added to the Grade Ten Examination in 1984. This was followed soon afterwards by the introduction of a continuous writing section into the Grade Six Examination, which serves as the entry qualification to secondary education.

Unfortunately there was little money to provide support for this initiative so that neither teaching material, nor help for teachers already in the field, were forthcoming to support the change in emphasis. The teachers did their best and, despite the difficulties, the students' writing at grade six and grade ten levels was seen to improve significantly in the years following the introduction of the writing examinations, a tribute to the washback effect. Writing skills improved, but after the initial spurt the level settled down to reveal that, like students everywhere, PNG students found academic writing difficult. The universities complained, as they had done before, that the schools were sending students who could not write, and continued to run their own pre-degree courses. But the problems of learning to write continued.

1.2.2 National syllabus

During the 1980's the teaching of writing in schools was supposed to follow a functional/notional syllabus. In practice, what happened was determined mostly by the requirements of the national examinations tempered somewhat by whatever materials were available for the teaching. The materials were further filtered through the perceptions of what teachers in the field considered appropriate and manageable. In the early 1980's multiple-choice examinations were used for English, but these did not encourage the development of writing skills. Students were able to write a tick or a cross, but not much

else. However, the introduction of continuous writing sections into the national examinations changed this, as mentioned above.

The kind of writing that was tested and therefore the kind of writing that was taught was functional English in the sense of being directly related to what the education officers and the teachers perceived to be relevant and possible. At grade six level (the final year of community school) it was considered realistic to expect students to be able to write a simple narrative of about half a page from a picture stimulus. Since the majority of students do not progress beyond grade six, this achievement fell terribly short of what educators would have liked students to be able to write at this level. At grade ten (the final year of secondary education for most students) the test requirements were for expository and argumentative or persuasive writing. Argumentative writing is referred to in my study as 'persuasive writing' since I prefer to describe the writing type according to its main writing purpose.

The prescribed curriculum meant that a student's secondary education tended to include very little personal narrative writing. Poetry or invented writing were even less likely to be practised because the type of writing practice tended to be dictated by the requirements of the Grade Ten Examination. Originally the examination included the writing of essays from graphs and tables, but since not only the students, but also the teachers found this task very difficult, this was rarely taught. The requirement for students to be able to write essays from graphs and tables was eventually dropped from the national examination in order to concentrate on those types of functional writing which seemed manageable.

1.2.3 Currently prescribed teaching of writing

From the brief description given above, it can be seen that educational requirements in PNG reflect an attempt to balance desired skill levels with what are conceived as realistic skill levels, and these are constantly readjusted to match changing educational beliefs as well as to take account of practical educational constraints. The most recent thinking of those who prescribe the curriculum has, in some ways, swung back to pre-1980 times to re-emphasise inventive writing as well as expressive writing. What is different in the current prescription is the view that the development of writing should be genre-based, i.e. that there is a hierarchy of writing types. The new language syllabi from preschool to

upper secondary (PNG Dept of Education 1995) have been constructed based on this view. The new syllabi are encountering the following practical problems:

- many teachers believe that errors should always be corrected and so this creates a marking problem if students are to be given a reasonable amount of writing practice;
- teachers (especially high school teachers) often question the relevance of creative writing.

1.2.4 Teachers' views on writing and accuracy

The PNG Department of Education believes that the practice of creative writing is a good way to help students develop their writing ability. In contrast, many teachers consider creative writing to be irrelevant and therefore a waste of precious time. Teachers often consider that expository and persuasive writing will benefit the students rather than creative writing. When there is limited time and a lot to teach, teachers believe that they should not waste time asking students to imagine that they met a green man last Wednesday. They think that the students would be better served by learning how to write, for example, a job application.

Teachers also worry about setting students frequent writing tasks of any kind because of the problem of finding time to mark them. Many teachers, and therefore also students, perceive the quality of written English to depend directly on an absence of error. It is an alien concept to disregard errors and comment solely, or mainly, on content. An essay that has not been corrected for mistakes is widely regarded to have been 'not marked'. This perception means that teachers are reluctant to ask their students to do much writing, because they do not have time to cope with the marking that would result.

The study attempted to address some of these problems. First of all it investigated, as part of the wider investigation into the relationship between the types of writing chosen for investigation, whether practice in writing imagined story narratives helped students to develop their persuasive writing skills. Secondly, as part of the investigation into which features of writing were associated with development and quality, the research tried to identify what role certain types of error played in the perception of writing quality and in the development of writing competence.

1.3 Aims of the research

1.3.1 Transition from narrative to persuasive writing

The main aim of the research was to find out how the transition occurred from an easy genre to a more difficult one, in this case from narrative to persuasive. In order to consider the difficulties and map the path of developing writing ability, it was necessary to examine the differences and similarities between different kinds of writing in order to explore how the transition was effected from one kind of writing to another. I decided to investigate the relationship between narrative writing and persuasive writing because narrative represents a kind of base line, which involves the kind of writing that is usually experienced as easiest. Narrative is the starting point of writing in school and persuasive writing is the desired end-point, demonstrating the achievement of an ability to express ideas and opinions logically and clearly. The problem to solve was how to help students move from one to the other, to find out what would help to increase competence in writing and especially in persuasive writing.

It was hypothesised that narrative writing fell into three categories because there appeared to be differences between them in the cognitive processes that were required for production. I labelled the three kinds of narrative writing other people's narrative (OPN), personal history narrative (PHN) and imagined story narrative (ISN), as shown in Table 2 below.

Table 2: Three types of narrative writing

	Other People's Narrative	Personal History Narrative	Imagined Story Narrative
Primary function	Enjoy, entertain & learn through retelling of stories of others	Maintain identity, validate self through retelling of personal experience	Play with new experiences, explore possible new situations, behaviour & consequences
Main cognitive processes	Memory selection	Memory selection, evaluation & ordering	Imagining a single or integrated event chain, requiring a recurring set of operations: choice, hypothesis & evaluation
- .....difficulty..... + - .....psychological interaction (inner dialogue)..... +			

Other people's narrative seems to be the easiest to produce, since it has already been given a structure by someone else. The primary functions are enjoyment, entertainment and learning through the retelling of someone else's story. The main cognitive process seems to be the requirement to select the story from the memory bank in order to retell it. This category includes myths and legends and any stories that have been heard or read or seen in films.

Personal history narrative seems to require more mental work in that the writer would need not only to select the experience from memory, but would need to order it and make sense of it so that it could be told as a story. The writer needs to function as a historian, and the cognitive processes needed for production require memory selection, contemplation and ordering so that the experience can be given a narrative structure and so that the writer can learn from the experience and make sense of it.

Imagined story narrative is different again in that it is necessary for the writer to choose an experience and possibly a different persona and then imagine an outcome. The function is to play with new experience through imagination and to explore it and learn from it. The writer is required to choose from an enormous range of possible experience and to imagine a path of possible consequences. This type of narrative shares with persuasive writing the need to imagine and to hypothesise, although it offers the comfort and security of being set in the past, albeit an imagined past. Once the imagined event chain is chosen, it becomes progressively fixed in a way that the possible consequences of ideas that need to be explored and compared in persuasive writing do not.

Since the research had to be limited, I decided to investigate practice in personal history narrative and imagined story narrative. These types appeared to require more cognitive effort than other people's narrative and I believed that practice in these would be more beneficial to the students than writing myths or legends or stories that they had already come across. A reason for choosing personal history narrative was that it was the narrative type with which the students were most familiar. It was already recognised as valuable writing experience and would hopefully provide a good contrast with any benefits to be gained from producing imagined story narrative in which students were likely to have

had little experience. Imagined story narrative appeared to form a bridge to persuasive writing in that part of the cognitive process required for its production was shared with the more difficult genre.

For these reasons, in order to describe the development of writing competence in high school students, I decided to investigate relationships between the three types of writing defined below:

Personal History Narrative (PHN) - telling about a series of events that has been experienced personally

Imagined Story Narrative (ISN) - telling about a series of events that has been invented by the writer

Persuasive Writing (PW) - expressing ideas and giving reasons in order to persuade the reader to agree

(Further explanation of choice of writing types is given in Chapter 3.)

1.3.2 Indicators of quality and development

A secondary and closely related question to the issue of how writing ability develops across genres, is the question of quality in a particular genre. What features of a piece of writing cause the writing to be judged as 'good' rather than 'poor'? Do these features hold over writing types? Do they remain the same as writing competence develops? Can 'good' writing be described and learned? How important is accuracy? Should language errors be given the importance they commonly assume in second language environments?

It was not possible to investigate the development of writing competence without considering the question of why a piece of writing was considered 'good'. The most common way of assessing quality is by subjective impression. We do this on a personal level when we read a novel, a poem or an academic article and it is still the most common method worldwide of assessing students' work. What we do not know is how the various elements of texts fit together to make them good pieces of writing, or why exactly perceptions of text quality vary. It is currently accepted that each reader creates his or her own text through the contribution of a particular culture and individual concerns (Cumming 1998, Raimes 1998), so that perceptions of writing quality can be expected to vary in ways that are hard to

predict. Consensus can be difficult to achieve, especially in regard to stages of writing development, where various writing skills are not developing evenly and a reader has to weigh up strengths in one area against weaknesses in another. Bearing this in mind, I tried to match holistic evaluations of scripts with some objective features of the texts to see which objective features were associated with a perception of quality. In addition, in order not to rely solely on subjective ratings, I assumed that writing development had taken place over three quarters of a year of writing practice and described some of the objective features that had changed.

It was not possible in one small piece of research to investigate all the elements of a text which can contribute to perceptions of its quality, so I decided to investigate the kind of features which teachers can easily recognise and can try to address. In particular it was considered important to investigate the contribution of fluency and accuracy to the perception of writing quality. Accuracy in writing was considered important since the subjects in question were writers of English as a second language. Teachers in general seem to focus heavily on 'mistakes' in writing and language teachers seem to do this more than teachers of other subjects. (See among others, Lukmani's 1993a, 1993b research findings on this issue, which are discussed in Chapter 2). Since the main research purpose was to try and describe the development of writing competence in order to enable teachers to be more effective, it seemed important to investigate the importance of the number and kinds of error on readers' assessment of text quality. The importance of fluency in writing was chosen for investigation in order to reach a view about whether to give more practice in order to encourage fluency, or more language teaching and correction in order to encourage accuracy. In addition to measures of accuracy and fluency, t-unit (minimal terminable unit)[1] measures were used to investigate differences in sentence structure because previous studies (for example Larsen-Freeman 1978 in an attempt to construct an ESL development index) had found t-unit measures to indicate writing development. (For discussion of t-unit studies see Chapter 2.)

[1] See Chapter 3.5.2 Table 8 for definition of 't-unit'.

1.3.3 Summary of aims

The aims were to find out which kind of narrative writing practice worked best to help students progress to persuasive writing and to investigate sentence structure, fluency and accuracy as indicators of development. The monitoring of the teaching and learning process during the writing project was a secondary concern because it was not possible to investigate all aspects of writing development in the same amount of detail.

The main aims of the research were the following:

1. to investigate the relationship between personal history narrative, imagined story narrative and persuasive writing as produced by grade nine high school students who speak English as a second language;
2. to chart the development of writing competence over three quarters of a school year in personal history narrative, imagined story narrative and persuasive writing;
3. to investigate the effect of practice in imagined story narrative, as opposed to the effect of practice in personal history narrative, on the development of writing competence and the transition to persuasive writing.

CHAPTER 2 - REVIEW OF LITERATURE ON THE DEVELOPMENT OF WRITING COMPETENCE

There has been a growing awareness that the ability to write is vital for personal, political and economic growth (Oxenham 1980; Martin 1985) and yet writing is almost universally perceived as difficult. The transition from narrative to persuasive writing, which is needed for academic and political success, is felt to be especially difficult (Britton 1975; Wilkinson 1983; Perera 1984, Yarapea 1990). As the English language increases its global dominance in economics, business and research, there is an ever increasing number of second language speakers who need to become competent writers in English. These learners bring with them an additional set of problems to the task of developing competence in writing (Johns 1991, Byrd & Nelson 1995).

Great claims have been made for the benefits of achieving competence in writing: faster learning and longer retention rates for literate people which lead to increased economic development (Oxenham 1980) as well as the promotion of higher forms of thinking (Vygotsky 1983). There is, however, no consensus on how writing competence may be achieved because there is as yet no definitive theory on either the development of language learning, or on the development of writing competence (Skehan 1988; Cumming 1998).

This chapter will report and evaluate i) the theories which seek to explain the process of writing development, ii) the factors which affect the development of writing, iii) ways of measuring writing and the associated problems, iv) the findings on objective indicators of writing development, and v) current pedagogy.

2.1 Theories of stages of growth in writing

2.1.1 Biological development and social need

The stages of growth in writing have usually been tied to stages of growth in thinking. Piaget (1972) saw the development of thought, reflected in speech and writing, as a series of intellectual operations linked to biological maturation. Although it is clear that physical maturation plays some part in the development of intellectual processes, it is Vygotsky's (1962, 1978, 1983) view that social need is the

major driving force behind the development of both thought and language. Luria, too, (1976, 1982) echoes the view of language development being driven by social need: that a person needs to think or speak in response to interaction with society and that when a person needs to think, say or write something then the means to do so will be found. The core feature of the process of thought and language is identified as interaction. Interaction is seen to take place initially with an external other, but after a while with an internalised other. It is Vygotsky's view that has gained favour (Hamp-Lyons 1991a), and there is an ever increasing emphasis in second language writing research on the importance and diversity of social context and social interaction (Cumming 1998; Kroll 1998; Raimes 1998). Included in the awareness of the importance of social context in writing is the notion of the social power of writing. Friere's (Freire & Macedo 1987) view of literacy as a form of cultural politics has been influential. Freire and Macedo believe that literacy becomes meaningful when it is viewed as 'a set of practices that functions to either empower or disempower people' (1987: xii). The acquisition of literacy that is driven by social need can, if it is an effective literacy, transform not only the individual, but also the society.

From the point of view of development of writing in the individual, Vygotsky (1983) points out that writing starts as second order symbolism where words are interpreted through the medium of sound e.g. children speak or subvocalise as they write. This gradually becomes first order symbolism as individuals begin to attach meanings directly to words without needing the intervening sound cues. It seems that as reading and writing develop, the sound link between the written word and its meaning becomes automatised. Readers and writers no longer need to subvocalise, but the auditory image remains important. Bradley and Bryant (1979 cited in De Goes & Martlew 1983) identified two major skills that are necessary in order to write: phonological skills i.e. the ability to link letters and groups of letters to sounds, and visual skills, where a learner learns to recognise whole words or sequences of letters. Liberman and Shankweiler (1979 cited in De Goes & Martlew 1983) note that reading and writing require a finer ability to analyse words into phonemes than does spoken language. Bradley and Bryant (1979 cited in De Goes & Martlew 1983) report a predominance of visual cues in reading, but phonological cues in writing. There is general consensus that auditory skills are crucial for written language (Snowling 1985), despite the fact that in all but beginning writers, the conscious need for

sound seems to have largely disappeared. The acquisition of the dual symbolism needed to cope with writing is a sophisticated mental feat. The difficulty is illustrated by the following comment by Nystrand:

> 'The writer's problem is how to say something with pen and paper despite the fact that not one single stroke, word or sentence corresponds naturally to what we call "thought"' (1982a: 58).

2.1.2 Increasing alienation of self

Following from Vygotsky's (1983) and Luria's (1982) view that the development of writing proceeds through interaction first with an externalised other and then with an internalised other, there is an increasing consensus that writing competence develops through a process of self-alienation. The idea that writing causes some kind of divorce between the writer and the self has been cogently and influentially expounded by Jacques Derrida (1978). Derrida argues that as soon as a writer writes, there is necessarily a split which means that what is communicated can never be 'true', that the act of writing engenders 'differance', a Derridean term meaning both 'to differ' and 'to defer' (Derrida 1978, Neel 1988). It is the act of selection, the picking out of part of a writer's truth from the rest of it that causes a split that alienates, and this is seen to happen more and more as writers attempt society-based genres, such as persuasive writing.

The consideration of 'differance' and the awareness of the increasing 'detachment' of a writer as s/he attempts society-based genres such as academic persuasive writing have occupied several writing theorists. Britton, in his study of first language speakers learning to write, comments that

> '..most children, learning to write in school, will proceed by dissociation, by a differentiation of performance, successively in face of differing demands.' (1975:114)

He, like Vygotsky (1983), believes that the process of writing consists of an internal dialogue between the writer and an 'internalised other'. Olson & Hildyard (1983) believe that the obvious divorce of the writer from the speech in academic writing is what gives the written word its authority. Geertz (1988 cited in Elbow 1991) emotively describes academic writing as "author evacuated" and Chafe (1985 cited in Brandt 1989) has written of the detachment of the writer. Deborah Brandt (1989) argues, however, that such detachment is an illusion and that the message in persuasive writing requires an

involvement equal to that required by more obviously self-involved kinds of writing. It seems that that the self is not detached when writing with an impersonal voice, but merely alienated and hidden.

The concept of the detachment or alienation of the writer through the act of writing extends Vygotsky's (1978) view that the inner dialogue between the self and an internal voice forms the basis of speech and language. Luria (1982) supported this view and emphasised that the function of the inner voice that we hear is not to resolve problems, because problem solving can occur much more quickly, but to regulate behaviour. The purpose of the interaction of the self with an internalised other and the degree to which the internalised other has to be distanced appears to be determined by the function of the kind of writing that is being produced. It is the extension of the idea of writing developing as an interaction with an internalised other that has given rise to the belief that the development of writing competence is associated with progressively difficult genres. As the genres become more difficult, the writer has to interact with an increasingly distanced other in order to produce successful text.

2.1.3 Hierarchy of genres

There is consensus that writing functions should be viewed as a hierarchy (Britton 1975; Freedman & Pringle 1980; Flower & Hayes 1981; Harris 1983; Wilkinson 1983; Perera 1984; Bereiter & Scardamalia 1987) and the methodology of genre analysis, which has arisen in response to the insight that it is the writing function that drives the process, attempts to show how the purpose of the writing is fulfilled through the text and to describe the differences between various types of writing (Swales 1990; Dudley-Evans 1994). However, there are varying characterisations of the order of the hierarchy and of the features associated with the various functions.

Moffett (1968) describes four discourse types, which he sees as a hierarchy starting with drama that has a recording function, progressing through narrative reporting, exposition and culminating in the theorising that requires logical argumentation.

Table 3: Moffett's hierarchy of discourse types

1.	what is happening - drama/recording
2.	what happened - narrative reporting
3.	what happens - exposition - generalising
4.	what may happen - logical argumentation, theorising.

Such a view of the writing type hierarchy seems at first to reflect a transition from personal speech to impersonal argumentation, a slide from the personal and specific to the impersonal and generalised. On reflection it seems that children or beginning writers find personal past narrative easier to produce than present commentary, yet Moffet's first two categories are in the opposite order. This may be because what has happened already has a structure, whereas what is presently happening is fluid and needs structure to be imposed as writing proceeds. Moffett's ordering of the specific to the general and the personal to the generalised, however, finds support in other hierarchies of writing types.

Britton (1975) builds on Moffett's model, but suggests a different organisation of the hierarchy. He constructed a hierarchy based on the belief that the development of writing proceeds from the expressive, which he sees as a kind of matrix from which differentiated forms of writing are evolved. He has at opposite ends of the continuum the 'transactional' (participant role - informative and conative) and the 'poetic' (spectator role - reflective), with personal expressive writing located in the middle of the continuum, and posited as its starting point.

Table 4: Britton's hierarchy of discourse types

participant role.....spectator role			
transactional		expressive	poetic
(more referential.....less referential)			
1.1 Informative record report generalised narrative/ descriptive info analogic speculative tautologic	1.2 Conative regulative persuasive		feelings, attitudes, beliefs of writer paramount poetic work must be taken as a whole, but a novel can be & usually is persuasive in that it has a message.

Britton's hierarchy is difficult to accept for several reasons and these are cogently argued by Perera (1984). Perera has three main criticisms. Firstly Britton's taxonomy is difficult to use from a practical point of view to classify texts, secondly there is no empirical evidence to support such ordering, and thirdly she points out that Britton's categories do not reveal anything about the organisational demands made on the writer. Perera (1984) suggests that a hierarchy of discourse types should contain a broad distinction between narrative and non-narrative writing on the basis that this distinction is the crucial one that separates information which is organised chronologically from information that is organised into a semantic hierarchy. Such a distinction gives rise to a hierarchy of writing types that seems more plausible. The past is easier to write about than the present and the future because, as noted above, the past already has a structure, whereas the present and the future require shape and hypothesis. Memory research supports such a view in its distinction between 'episodic memory' where events are ordered chronologically, and 'semantic memory' where they form a meaning network (Gregg 1986).

Bereiter and Scardamalia (1987) make a distinction along similar lines when they distinguish between 'knowledge telling' (a narrative style of operations) and 'knowledge transforming' (an expository/ persuasive semantic hierarchy type of operation), although they focus on stages of competence in writing rather than on a hierarchy of writing types. The same distinction between past and present/future is incorporated in Phillip's (1995) hierarchy of discourse types.

Table 5: Phillip's hierarchy of discourse types

past.....future
- .....psychological distance..... +
- .....audience ..... +

voices:	historian	reporter	storyteller	evaluator	expositor	persuader
main	1	2	3	4	5	6
function:	tell own experience (to understand)	tell society's experience (to learn)	tell invented experience (to play)	evaluate own experience (to reflect)	explain society's experience (to predict)	evaluate possible experience (to persuade)
time:	past	past	past	present	present	present/future
	my past world	*our past world*	*our possible past world*	*my present world*	*our present world*	*our possible world*

- .....difficulty..... +

The bottom end of the hierarchy runs from writing set in past time, which tells about various kinds of experience in order to observe and reflect, to writing set in the present and future, which evaluates various kinds of experience and ideas.

There are empirical findings to support the view of a hierarchy of difficulty between genres. Perera (1984) in her study of first language writers noted the general distinction of difficulty between narrative and non-narrative writing, where narrative was easier. The view that writers have difficulty making the transition from narrative writing to academic types of writing where explanations, comparisons and hypotheses are needed was supported by Yarapawa (1991). He reported that at the University of Technology in Papua New Guinea students could write recounts, procedures and reports, but could not write explanations or effective discussions.

There seems, however, to be very little research on the relative difficulty of producing different kinds of narrative text. Van Bruggen (1946) compared the production of two types of narrative writing. He compared L1 students' retelling of a story they had heard with their production of personal narrative accounts, i.e. a comparison of OPN with PHN. He found that students produced recounts of the story

they had heard much faster than they produced their own personal narrative accounts. This supports the speculation that the OPN category is easier to produce than PHN, presumably because it already has a form. More recently, a study comparing the development of oral narrative in monolingual and bilingual children showed that there was a difference between them, where bilingual children showed a comparative lack in the kind of text features, such as evaluation, that could be said to contribute towards academic literacy (Shrubsall 1997). Although Shrubsall did not compare different kinds of narrative, the guided ISN type of writing used for the study, where children were given pictures with no text and asked to make up a story, illustrated the hypothesis and evaluation needed for this kind of narrative and noted its usefulness in preparing children to move on to more academic forms of discourse. It was argued that the findings revealed part of the academic gap, theorised by Cummins (1984 cited in Shrubsall 1997), that bilingual children still had to cross in order to perform well in school. The same argument could apply to writers of English as a second language.

That students experience greater difficulty with persuasive writing than with narrative writing (Freedman & Pringle 1980; Hamp-Lyons 1991a), is partly explained by the difficulty of the more distant audience considerations that are needed for persuasive essays. Many researchers draw attention to the difficulty of anticipating and taking account of the readers of a text, particularly for beginning writers (Kroll 1978 cited in Cooper & Matsuhashi 1983; Flower 1979; Bereiter & Scardamalia 1981; Britton 1983; Martlew 1983; Watson 1983; Intaraprawat & Steffensen 1995). The difficulty imposed by more distant audience requirements appears to alter the structure of the language to make it more formal. Crowhurst & Piche (1979) found that when a more distant audience was specified, that tenth graders produced longer t-units[1] and longer clauses.

Purpose or type of discourse has been found to consistently correlate with syntactic complexity (Watson 1983) and Rosen (1968 cited in Britton 1975) commented that more able writers showed the biggest t-unit differences between genres. When syntax is more complex, it is more difficult to produce and Bartholomae (1980) noted that a high percentage of error was rooted in difficulty of

[1]See Chapter 3.5.2 Table 8 for definition of 't-unit'.

performance rather than in linguistic competence. This view was supported by Peterson (1993), when she found that children made more errors when they wrote than when they spoke.

Finally, a point to bear in mind in the consideration of the levels of difficulty of various writing types, is the view that writing types are never discrete. Mlynarczyk (1991) notes that there is often an overlap of genres, where the overall writing purpose leads to a classification that ignores the presence of other types of writing within it. Reppen (1995) observed the same phenomenon and noted that students, particularly when inexperienced in academic writing, repeatedly tried to turn their academic writing into stories.

In summary, it seems that the growing consensus on a hierarchy of discourse types stresses a separation in terms of difficulty between narrative structures dealing with the past and the structures of academic types of writing that deal with present and future possibilities. Hierarchies of discourse types need to take account, too, of differences in audience perception and the formality of the type of writing. It is worth remembering that the level of difficulty involved in producing a particular piece of writing must result from a complex interaction of many factors that combine to determine the amount of motivation felt by the writer. Such interactions are not yet fully understood.

2.1.4 Generative and recursive writing process

The process of writing is not only self-alienating as discussed in 2.1.2 , and increasingly so as more difficult genres are produced, as discussed in 2.1.3, but it is also generative and recursive. The stages of growth in writing generate stages of growth in learning. It is the view of numerous researchers that we find out what we think through writing, and as we write, new thoughts are generated in a process that is endless (Derrida 1978; Flower & Hayes 1981; Zamel 1982, 1983a; Britton 1983; Scardamalia & Bereiter 1983; Raimes 1987; Boughey 1997). Johns (1990) believes that the writing process is a creative act where the discovery that results is more important than the product. Luria comments that

> '..it is not understanding that generates the act, but far more the act that gives birth to understanding - indeed the act often far precedes understanding'
> (1983: 276)

The process is also recognised to be recursive (Flower & Hayes 1981; Scardamalia & Bereiter 1983; Cooper & Matsuhashi 1983), so that writers are believed to switch to and fro in a non-serial manner between planning, creating text and revising. The generative, recursive nature of the writing process was clearly documented in Emig's (1971 cited in Zamel 1982) influential work on the composing processes of twelfth graders. She found that her subjects displayed varying behaviours and that in none of them could the composing process be said to conform to a set of discrete stages. Since then Flower and Hayes (1981), Bereiter & Scardamalia (1983), Zamel (1983a) and many other researchers have reached the same conclusion. The recursive nature of the writing process seems to intensify as writers become more advanced. Good writers appear to produce more drafts with more substantial changes than do poor writers (Zamel 1982; Bereiter & Scardamalia 1983; Martlew 1983; Graves 1984; Fitzgerald 1987; Connor & Asenavage 1994). This seems to imply that the involvement with a piece of writing lasts longer as writers develop their competence, although the writing purpose will obviously affect the number of drafts and the time taken over a particular piece.

2.1.5 Summary

The development of writing competence seems to be driven by social need and to proceed through the interaction of self with other to produce text. The process of composing is self-alienating, generative and recursive. These three features of the writing process appear to increase in intensity as writing competence develops and as writers attempt increasingly difficult genres. The various factors that affect the development of writing competence will be discussed next.

2.2 Factors affecting the development of writing

2.2.1 Maturation

The development of intellectual skills was described by Piaget (1972) as closely tied to physical maturation. In contrast, Vygotsky (1983) and Luria (1976), although accepting physical maturation as a precondition, believed that development was powered by social need and could occur at any time. Although it is the latter view that has gained favour, as reported above, it was Piaget's view of the link between maturation and intellectual development that had, initially, the greatest influence on research into the development of writing and attempts to measure it.

The development of the syntactic complexity needed for writing is considered by Hunt (1983) to be closely associated with physical maturation. He claimed that sentence length increased with age and that the number of words per t-unit (minimal terminable unit), as well as the number of clauses per t-unit and the number of words per clause, increased with age because of the increase of noun modifiers and increased nominalisations. This observation clearly assumes schooling and a progression from narrative to an academic type of writing. Young (1985) commented that the development of syntactic maturity in writing could be seen as a process of gradual control of the marked structures of the language, and elaborated further to explain that marked structures contained an element of meaning absent from unmarked structures. It is from the work of Hunt and his followers that we received the now widely accepted term 'syntactic maturity' and by extension 'writing maturity'.

It is clear that the structure of proficiency changes as writing competence develops (Bachman & Palmer 1981a, 1981b cited in Skehan 1988), although the development of competence is not necessarily related to physical maturation. Witte (1983) emphasised that t-unit length stabilised only after repeated practice of writing skills. The emphasis on maturation as a strongly determining factor in the development of writing competence is due partly to the fact that research into the development of writing competence was undertaken first in relation to L1 writers who were studied as they progressed through school. The concept of 'syntactic maturity' became generalised to mean the syntax of advanced writers, and has been used by research into second language writing development. It was employed by Larsen-Freeman (1978) in an attempt to construct an ESL index of writing development, on the basis of a previous finding that t-unit length and the number of error-free t-units were the best measures (Larsen-Freeman & Strom 1977). In the second study, Larsen-Freeman (1978) found that the percentage and length of error-free t-units were the best objective discriminators among five levels of writing proficiency.

Although it seems sensible to accept that physical maturation affects stages of growth in the development of writing competence, it is important to note that physical development is not the only factor affecting the growth of cognitive processes. Hamp-Lyons (1991a), following Luria (1976) and

Vygotsky (1983), makes the point that cognitive development and hence also the development of writing competence, is not something that occurs only in school, but can occur throughout life.

2.2.2 Schooling

It used to be the view that the acquisition of literacy in itself provided the crucial benefit in order to help a person develop (Oxenham 1980). Scribner and Cole (1981), however, in their work on the psychology of literacy, came to the conclusion that the significant increase in reasoning powers came from the effect of schooling, rather than from the effect of literacy alone. They were able to investigate this issue by studying the Vai people in Liberia, who acquire basic literacy skills, even when they do not go to school. It is necessary to remember when considering their findings, however, that the Vai people's literacy, as reported by Scribner and Cole, was a basic literacy consisting of formulaic letter reading and writing. It was not the type of writing which required exploration of thought, or invention of ideas. Scribner & Cole cited research carried out by Greenfield & Bruner (1966 cited in Scribner & Cole 1981) on concept formation of Wolof children in Senegal, where once again the conclusion was that the differences in the groups studied were attributable to the effect of schooling rather than to the effect of literacy alone. It is worth noting though that the effect of schooling contains within it the effect of literacy and the findings of Havelock (1963 cited in Scribner & Cole 1981) concluded that literacy was a precondition for the emergence of universal concepts. Scribner & Cole's own research (1981) found that while schooling did not increase ability, it did have the effect of enabling students to explain their performance. In other words, schooling increased their reasoning powers. Scribner and Cole also found that another effect of schooling, which included the effect of literacy, was to increase memory.

Scribner and Cole's (1981) research shows that the development of writing competence depends not only on the acquisition of literacy, but also on the contribution of schooling. It shows the importance of teaching and the stimulation of thought that results from it. It should be noted, too, that the kind of schooling received can vary enormously both between individuals and between cultures. Carson (1992) emphasises that the social context of schooling and the pedagogical practices most often used

for teaching reading and writing have obvious implications for the acquisition of second language writing skills.

2.2.3 Knowledge of text schema

Stubbs (1982), when speaking of first language speakers learning to write, emphasised the importance of the understanding children need of the relationship between the spoken and the written language in order for them to be effective learners of how to write. Children need to be aware that written language is different from spoken language, and they need to acquire the schema associated with different types of written text in order both to understand and to produce a piece of writing that belongs to a particular genre (Rumelhart 1981). It is the power of expectation that derives from the knowledge of text schema that is crucial both for understanding and for producing text (Tannen 1979 cited in Cummins 1983). For immature writers, who are learning to write in their first language, the main difficulty can be a lack of knowledge of written texts (Perera 1984). Students have difficulty in producing school type prose because 'home-based discourse strategies differ from those of the school'(Collins & Michaels 1986 cited in Brandt 1989:33).

There is a consensus that knowledge of text organisation, i.e. the schema of a text, is important for writers (Rumelhart 1981; Britton 1983) and the positive effect on L2 writing of book flood projects (Elley 1994 cited in Cumming 1998) supports this view. The 1956 UNESCO report, however, quoted by Oxenham (1980) states that all good writers are good readers, but that not all good readers are good writers. A knowledge of rhetorical pattern would seem, then, to be a necessary, but not a sufficient condition for writing.

2.2.4 Writing in a second language

Although the development of writing competence is considered to be similar in many ways for all writers (Widdowson 1983; Silva 1993), it is clear that second language writers are likely to experience constraints caused by their more restricted knowledge of the language and the culture it reflects (Johns 1986; Frankenberg-Garcia 1990; Ballard & Clanchy 1991; Santos 1992; Dyer 1996). Devine, Railey, and Boshoff (1993), in a comparison of L1 basic writers and L2 writers, found that L2 writers did not

hold the same cognitive models as the L1 writers, although the L2 students differed only in their increased focus on accuracy. Flower and Hayes (1981) identify three constraints on the composing process:

1. knowledge, which they admit should normally be seen as a resource, but which they claim is a restraint when it is inadequate;
2. written speech (because it is different from spoken speech), and
3. rhetorical pattern.

For writers in a second language the knowledge of the target language may be so deficient that their ability to produce written texts is seriously hindered. Insufficient knowledge of language will increase the load on Short Term Memory (STM) for second language writers compared with first language writers (Flower & Hayes 1981; Cooper & Matsuhashi 1983), and may lengthen the process of text production because of the need to revise language more intensively (Zamel 1983a), and more frequently (Silva 1993). The added cognitive difficulties faced by ESL writers are emphasised by Flower and Hayes (1981) when they describe the process of writing as a dynamic act where a large number of demands or constraints have to be dealt with simultaneously. They make the point that for many ESL writers, the demands on STM may be excessive.

A specific area of language deficiency common to many ESL writers has been identified as difficulty with linking words (Bacha & Hanania 1980; Zamel 1983b; McDevitt 1989; Demel 1991). Knowledge of linking words and the ability to use them correctly is clearly important for the production of coherent text. Production of coherent text depends heavily on knowledge of the rhetorical patterns associated with the text schema of a particular genre. Second language writers can be disadvantaged in this area if they have had insufficient exposure to written texts in English. They may also be disadvantaged by a natural reversion to the cultural norms of rhetorical organisation that are used in their first language.

The influence of cultural views on performance in academic writing is emphasised by Ballard and Clanchy (1991), since cultural differences in the organisation and presentation of information may

conflict with what is conventionally expected in the target language. Basham & Kwachka (1991), for example, found that Alaskan students, writing summaries in English, showed a tendency to include their own views or even personal narration. It is worth observing, however, that writers in general, both L1 and L2, may have a tendency to mix writing types, e.g. narrative and persuasive, as their competence develops. It may be that Basham and Kwachka were observing a mixing of writing types in beginning writers rather than a culturally-specific text organisation.

Cultural familiarity helps text comprehension (Carrell 1983, 1985, 1987) so it would seem sensible to expect that the cultural familiarity of text organisation as well as text content might be helpful in the composing process. Winfield and Barnes-Felfeli (1982) lend support to such a view in their research on the effects of familiar and unfamiliar cultural context on foreign language composition. They found that familiar cultural context aided written recall, which was manifested as an increase in fluency, an increase in grammaticality, and a lack of inappropriate own culture interpretations. It could be argued, however, that their research had more to say about text comprehension than about text production. Silva (1993) reviewed studies that had investigated L1/L2 differences in text production and reported that three studies had found strong similarities in patterns of logical relations used in texts, although other studies had found various differences in rhetorical organisation that could, presumably, be traced back to the mother-culture of the L2 writers. Yu and Atkinson (1988 cited in Silva 1993) found that L2 writers linked arguments less effectively than L1 writers, but this seems more likely to have been due to a restricted knowledge of the target language than to a conflict between cultural differences in text organisation.

From the discussion above (2.2.3) it is clear that knowledge of text schema is an important precondition for the production of text, but a greater familiarity with a rhetorical organisation different from that of English may operate as a constraint on the composing process of second language writers. The differing rhetorical systems of the first language were claimed by Kaplan (1966, 1967) to cause a substantial difference in the composing process of L1 and L2 writers. Later on Kaplan (1987 cited in Hamp-Lyons 1991a) softened his position slightly to incorporate a broader view of content and schema, but still believed that rhetorical systems differed and that this constituted a significant

composing constraint on second language writers. The importance of rhetorical style, which differs according to the target audience, even for first language writers, is emphasised by Regent (1985) and there is substantial support for Kaplan's main point that cultural considerations constrain the production of text in a second language (Kobayashi 1984 cited in Hamp-Lyons 1991a; Ballard & Clanchy 1991; Basham & Kwachka 1991; Reid 1992). Zamel (1983b) disagrees. She believes that L2 writers have additional composing difficulties because of the level of their knowledge of the language and particularly because of their inadequate understanding of the meaning of linking words and how to use them, rather than because of a need to write using a different rhetorical system. Bacha and Hanania (1980) share her view.

The view that the development of writing competence is essentially the same for both first and second language writers is held by several researchers who make the point that composing skills and language proficiency are not the same (Widdowson 1983, Zamel 1983a). Similarities in the probable composing processes of first and second language writers are emphasised by Aitchison (1987). She notes that human beings need to know three things about words in order to be able to use them effectively: their meaning, their role in the sentence, and what the word sounds like. Evidence from Green (1986 cited in Aitchison 1987) and other researchers points to the likelihood of a bilingual word store implying a single integrated network of words that writers choose from. Such findings do not diminish the fact that inadequate knowledge of a language hinders both writing and speaking, but does indicate a similarity of composing processes, in that all writers need to operate similar processes and that all writers experience a similar relationship between the spoken and the written language.

The difficulties of transforming spoken language, or the language of thought, into written language are emphasised by numerous researchers (John-Steiner & Tatter 1983; Ali 1989). Mohan and Au Yeung Lo (1985 cited in Ali 1989) reported that L1 and L2 writers had similar difficulties in this respect. The differences between spoken and written language are well documented, particularly by Halliday (1985) who considers that spoken and written language have different functions. Although both functions can be expressed by both spoken and written language, Halliday sees the main purpose of spoken language as pragmatic (for doing) in contrast to the main purpose of written language, which he identifies as

mathetic (for learning). Halliday also points out differences in form between spoken and written language where 'the complexity of written language is lexical, while that of spoken language is grammatical' (1985:63). An example of this difference is given by Altenberg (1984 cited in Ali 1989) when he points out that spoken language relies on 'but' and 'so' to link ideas, whereas written language relies more on lexical and structural variations. Bartholomae (1980) argues that L1 speakers have to learn the equivalent of a second language when they learn to write. The effort of converting spoken into written language applies to writers regardless of the language they are using to express themselves.

Given the common sense assumption that a certain threshold of skilled language use must be a precondition for writing, it seems that most researchers believe the composing process of second language speakers to be broadly the same as that of first language speakers (Zamel 1982, 1983a; Krashen 1984; Arndt 1987; Raimes 1987; Uzawa 1996; Kamimura 1997; Cumming 1998). Widdowson (1983) emphasises the point when he observes that the writing difficulties of second language writers are certainly not linguistic in a straightforward way, since their composing difficulties are not solved merely by acquisition of language competence. Raimes (1987) maintained that, in the sample of ESL college writers she studied, there was little correspondence between language proficiencies, writing abilities and composing strategies, but Kamimura (1997) found that the L1 and L2 writing abilities of her Japanese students were positively correlated once a certain threshold language level was passed. This view was supported by Hirose and Sasaki (1994), who found that the L2 writing ability of their Japanese students, as revealed by samples of expository writing, was correlated with L2 language proficiency and L1 writing ability. It seems that reading skills transfer more easily between L1 and L2 than do writing skills (Eisterhold-Carson J., P. Carrell, S. Silberstein, B. Kroll & P. Kuehn 1990) and that reading skills in a particular language are usually acquired before writing skills.

In summary the development of writing competence for L2 writers can be hampered by additional difficulties not faced by their L1 counterparts. The most significant additional difficulty is an imperfect knowledge of the language in which they are writing. The main point to make is that L2 writers do not

always acquire the different components of written control at the same rates (Hamp-Lyons 1991b). For example, it is possible for writers to acquire fluency without accuracy or accuracy without fluency. First language writers may have similar problems for other reasons. For example they may have poor spelling because of insufficient exposure to the written language, or the load on STM required for a particular piece of writing may compromise the accuracy of their sentence construction, as well as their overall text organisation. It is possible, too, to have a wide vocabulary but poor syntax, or alternatively syntactic control but no rhetorical control (de Jong & Henning 1991 cited in Hamp-Lyons 1991b). Pollitt & Hutchinson (1987) make the same point in a different way, when they note that the difficulty of engaging with a task is task-specific and that the effective achievement of a task depends on competence in the components of writing skills needed for that task. The fact that second language writers are subject to the same constraints as first language writers in addition to the added difficulties of composing in a language with which they are not totally familiar either syntactically or culturally, means that it is not easy to be sure of the reasons for second language writing difficulties. The problems are evident, but the reasons may not be.

2.2.5 Motivation

A writer's motivation is influenced by the degree of difficulty imposed by the writing, by the level of involvement with the topic, and by his or her attitude to the reader/s for whom the text is produced. Many researchers comment that writing is difficult and this is the most obvious constraint on the act of composing. Widdowson (1983), for example, believes that the result is not equal to the effort of writing. The level of difficulty imposed by a piece of writing is the result of the interaction of various factors operating on the writer, but some of the difficulty results from the kind of writing that is required and the amount of cognitive load it exerts.

Persuasive writing, for example, exerts a greater cognitive load on STM than narrative writing because it is more formal and has a very different structure from that used in narrative, which is closer to spoken language (Halliday 1985). The nominalisation and subordination of the text structure required for persuasive writing require long stretches of text to be held in STM because clauses and their referents are not necessarily adjacent. Pieces of text have to be held in memory until their referent can

be identified and the meaning of the sentence processed. In addition, subjects and the verbs they govern are not necessarily adjacent either and similar constraints apply. With narrative writing, although the sentences may be long, the structure is simpler because units of meaning are usually ordered serially.

Not only is the internal structure of a typical persuasive writing sentence more difficult to process and produce than a typical narrative sentence, but the overall text structure is harder to hold in memory in order to continue with the writing. Thinking does not end with the answer, it is followed by evaluation (Luria 1973), so the meaning of the preceding text has to be accessed in order to evaluate. Persuasive writing usually requires the STM to hold a comparison of ideas and probabilities that relate to each other in increasingly complex ways as additional information is added in the developing text, whereas narrative writing normally requires the memory to hold just a single event chain. (This is a slight simplification, since narrative actually has a partially, not a totally, specified trajectory in episodic memory base (Golden & Rumelhart 1993), but the comparison stands.)

Memory research confirms that the five to seven items of information that can normally be held in STM can appear to be multiplied if those items link clearly and easily to other pieces of information (Gregg 1986). This is how mnemonists perform apparently impossible feats of memory: by linking an item to be remembered to another item, which is linked to another, and so on. Narrative writing usually requires the writer to run back along only one path, which means that the latest item of information can be used to pull out the rest. Because of the complexity of the relationships between the information expressed in persuasive writing, the task of holding in awareness what has gone before becomes much more difficult. Motivation is obviously affected by how difficult the writing is perceived to be. A heavy load of information for the brain to manipulate is perceived as unpleasant or even painful to the extent that the writer feels a great reluctance to write and will do almost anything in preference to producing the required text.

The level of difficulty of the writing may be the most obvious constraint on the writer, but it may not be the most powerful one. The degree of interest and involvement generated by the topic is a crucial

motivating force, since this is what drives the writer to write and keeps the writer going to overcome all difficulties. Watson (1983) stresses the need to consider the importance of topic as a variable in affecting students' achievement of a writing task, and yet the kind of writing students are required to do in schools often does not correspond to their interests or needs (Stubbs 1980 cited in Gundlach 1982; Bruton 1981). If a writer has no desire to say anything on the prescribed topic, s/he will find it very difficult to write.

The writing task needs to be not only interesting, but also clear. The student needs to know exactly what is being required so the writing goal can be clearly formulated (McKay 1980; Watson 1983, Arndt 1987; Yarapawa 1991; Reid & Kroll 1995). Yarupawa (1994) comments that students in the University of Technology in Papua New Guinea were frequently unclear about what exactly a particular writing task required and that they underperformed accordingly. Horowitz, too, (1989, 1991) stresses how important it is that the writing task should be clearly understood by the students.

The interest generated by the topic and the clarity of the task make it easier for the writer to form a clear goal or writing purpose. Flower and Hayes (1981) comment that writers need to create two kinds of goals - an overall goal that has to be kept in mind and sub-goals which enable the developing sense of overall purpose. They also comment that goals may change during the act of writing and it is at this point that a writer may lose the way. There seems to be a consensus that the flexibility to change goals during the writing process is necessary for effective execution of text and this is supported by many researchers, such as Britton (1975, 1983), and Rose (1984) in his study of writer's block. The changing of the writing goal implies the need for possibly extensive revisions and to do this the topic needs to continue to generate a high level of interest and the writing task needs to be kept clear. If the topic is not sufficiently interesting then the writer may feel that the rewriting is not worth the effort.

The writer's relationship with the intended audience affects motivation. Since different types of writing impose different audience requirements, the type of writing affects the writer's relationship with the reader. Narrative writing, for example, expects a familiar close audience, while academic writing expects a formal, distant audience. The emotional impact of the writer's attitude to the

potential readers can work either positively or negatively. Horning (1993) discusses the effect of the writer's emotion on the production of text, and notes how powerful this can be. Brandt identifies audience considerations as a significant composing constraint. She comments that:

> '...we have not done justice to the writer's attachments to readers that evolve as a text develops. The demands for what we call coherence, consistency and even clarity are really commitments to an evolving relationship with those we are sharing a text with - a recognition of the presence of the other. Such attachments become an overriding constraint during composing.' (1989:37)

The kind of feedback on writing a student has previously received from the teacher contributes to the positive or negative perception of audience and consequently affects the motivation to write. Unfortunately, teachers, and particularly teachers of English as a second language, often emphasise the evaluation and correction of the expression more than they respond to text as a communicative act (Gundlach 1982; Reid & Kroll 1995). Negative feedback can seriously damage the self-confidence of the writer.

Second language writers are affected by their attitude to the speakers of the target language. The attitude can be positive or negative, but in many countries with large groups of ESL learners there may be a negative memory of colonisation. Suspicion or dislike of the target language speakers may reduce motivation to learn or write in English. And yet, for ESL learners, unlike students of EFL, the acquisition of skills in the target language is often necessary for survival . Second language writers are affected not only by their attitude to the language being learned, but by the amount of anxiety or the amount of self confidence they bring to the act of writing (Gardner & McIntyre 1992, 1993; Hirose & Sasaki 1994). Negative emotion associated with the writing can heavily reduce the urge to keep trying.

2.3 Subjective evaluation of text

Writing is not only difficult to do, it is difficult to measure because of its multifaceted nature and the fact that text is seen as the midpoint in the social interaction between writer and reader/s (Vygotsky 1962, 1978; Luria 1982; Nystrand 1982b; Prucha 1983; Rocklin 1991; Hamp-Lyons 1991a). The problem underlying all measurement of writing is that there is no definitive theory of writing, either of what exactly constitutes quality, nor of how competence in writing develops. In a discussion of

theoretical models, assessment frameworks and test construction, Chalhoub-Deville (1997) concludes that the consensus favours a componential approach to language proficiency but that the nature of the components remains debatable, a point also made by Choi (1992). It is noted that some models are questionable either because they are not based on empirical findings, or because they are based on incorrect or incomplete findings. The ELTS (English Language Testing System) test, which later became the IELTS (International English Language Testing System) test, is based on the view that language proficiency is divisible, not unitary (Alderson & Clapham 1992). This is the common view, but it does not help very much because the effects of one feature on others are so little understood. Skehan (1989) states that it would be helpful to see measurement related to developmental stages in language learning, but that so far testing tends to be dictated by writing syllabuses, themselves subjective views. He comments that the testing of language, including the testing of writing, has been 'limited by the deficiencies of the syllabuses on which they [the tests] are based.' (1989:9).

As an indication of how inadequate our present testing is, we have only to look at Geranpayeh's (1994) study, which compared two of the most widely recognised and prestigious tests: IELTS and TOEFL. He found that although score comparisons were more or less justified, subjects with similar language proficiencies did not score the same on both tests. Hamilton, Lopes, McNamara and Sheridan (1993) compared NS and NNS performance on various internationally recognised tests. They reported that the TOEFL test did not test performance but tested knowledge and that NSs did better on the test than NNSs, in contrast to communicative tests such as TEEP (Test of English for Educational Purposes) and IELTS, where NSs often did not do better than NNSs, particularly in reading. In these kinds of tests, what seemed to matter most was the level of education and the educated scored better than the less educated, in contrast to the TOEFL test. Hamilton, Lopes, McNamara and Sheridan draw attention to the fact that the NS has been used as a reference for NNS testing since the 1950's, especially in terms of Chomsky's 'ideal native speaker/hearer'. They question the usefulness of this, especially in terms of the differences between oral and written language. The point is that although a native speaker can produce oral language with ease, it does not follow that s/he can produce effective written language just because s/he is a native speaker of the language. There is also the question of different varieties of English. Leung, Harris and Rampton (1997) question the notion of the ideal

native-speaker/hearer and recommend that teachers should be more concerned with questions of 'language expertise, language inheritance and language affiliation' (1997:543).

Until the 1980's, writing quality was frequently assessed by objective measures of skills that were felt to be related to writing, for example cloze tests and some multiple choice tests, which seemed to correlate with direct writing assessments. Tests like these were used because they were shown to be highly reliable, but they had poor face validity. Teachers felt that ability to produce good writing should be judged by the actual production of writing and not by related skills. Face validity considerations were not in any way minor ones, because it was primarily teachers' opinion on the form of the TOEFL test that forced the change from objective testing to a direct test, i.e. the Test of Written English (Greenberg 1986). Hamp-Lyons (1991c), however, emphasised that the most powerful argument in favour of direct testing was its washback effect.

Doubts about the reliability of subjective testing still remained, although research such as that carried out by Kaczmarek (1980), which found that subjective tests worked as well as objective tests, helped to allay fears about unreliability. The consensus was, and still is, that holistic evaluations in one form or another are best because the skill of writing is believed to be made up of various components which somehow need to be tested even though their interrelationships are not properly understood (Huot 1990; Chalhoub-Deville 1997). Such a view arises partly because of the poor face validity of indirect tests of writing and partly because of the awareness that until a full description of what makes a good piece of writing is available, our intuitive impressions are presumably more accurate than incomplete theories of text. Cumming (1990), for example, makes the point that writing proficiency is distinguished from language proficiency in holistic ratings, and we can assume that intuitive scoring accesses and computes all the myriad features of text and their effects on one another in ways that objective analyses cannot.

There now follows a discussion of methods of subjective evaluation, problems with rating and task variability.

2.3.1 Methods

Analytical scale scoring was developed in the hope of achieving greater inter-rater reliability and was made popular by the Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey (1981) Composition Profile. The idea was to score holistically while bearing in mind particular preset criteria. Although it may have been effective in improving inter-rater reliability and in providing more detailed feedback for students where this was appropriate, there were arguments against it. The most powerful argument against analytical scoring was made by Mathews when she asserted that the testing of subskills separately was invalid because '...it is illogical to allocate equal marks to the various sub-skills as though the relationship between them was a simple one of addition' (1990:118). While analytical scales may not always allocate equal numbers of marks for each subskill, they do allocate an arbitrary number of marks, so Mathew's point has to be taken into account. The second criticism was that forcing raters to conform to preset criteria may compromise the validity of their evaluations.

Another variation on holistic scoring with an analytical scale was primary trait scoring, where a particular writing task had a special set of scoring criteria devised for it. This method was recommended (Cooper 1977 cited in Ali 1989, Odell 1981 cited in Ali 1989) on the grounds that the rater's attention should be drawn only to features that were relevant for the fulfilment of the particular task. A criticism of primary trait scoring was the doubt about whether it was possible for a reader to ignore all other facets of text quality and judge only one thing (Hamp-Lyons 1991b). A development of primary trait scoring is 'performative assessment', which is similar except that 'performative assessment differentiates and assesses skills that are all lumped together in primary trait assessment' (Allaei & Connor 1991: 228). Henning (1991) and Hamp-Lyons (1991b, 1995) recommend 'multiple trait assessment', which is similar to performative assessment in that it 'implies scoring any single essay on more than one facet or trait exhibited by the text' (Hamp-Lyons 1991b: 247). The main consideration is that the act of writing contains many different skills that do not all develop uniformly, so that a single score will not identify strengths and weaknesses to inform teaching or learning needs (Carroll 1983, Hughes 1989). Multi-trait assessment is recommended because the method ensures feedback to students in specific skill areas (Henning 1991; Hamp-Lyons 1991b, 1995). It would, however, be very expensive to provide feedback for students in mass testing situations where the

purpose is to evaluate writing proficiency levels for entry into courses of various kinds. Perkins (1983) makes the point that different tests are suitable for different purposes, i.e. holistic tests have the highest construct validity, while analytical tests are best for classrooms because they diagnose problems.

One problem of holistic scoring is that differences between subjects can be small and evaluators find it difficult to discriminate between small differences (Upshur & Turner 1995). The problem applies whether or not an analytical scale is used to componentialise the assessment because the final rating for the components is still holistic and the band remains broad. Polio (1997) commented that the holistic scale she used in her study was problematic because the inter-rater reliability was low and yet the raters felt that the scale could not be modified to make it any more reliable. They felt that the scale could not be constructed so as to distinguish differences in linguistic accuracy. A second problem is that despite descriptors, raters often cannot help but norm-reference (Carroll 1982 cited in Mathews 1990, Sheridan 1991 cited in Hamilton, Lopes, McNamara & Sheridan 1993). The two major criticisms of holistic scoring are that the rating is highly subjective and that students perform differently on different topics (Kaczmarek 1980). Rating problems will be discussed first.

2.3.2 Rating

To rate requires reading and features of text, such as handwriting, spacing, syntactic, semantic, textual and contextual aspects, influence readers in personal ways (Nystrand 1982a). This means that there is bound to be variability between raters as there is between readers (Purves 1992). Vaughan (1991) stresses this point and concludes that cursory reading of a single writing sample by two essay raters should not be a basis for passing or failing, as it so often is.

Doubts about inter-rater reliability are a source of concern to all those who are engaged in the testing of writing. Holistic impression marking is widely regarded as the method of writing evaluation with the highest construct validity (Perkins 1983), but the literature offers conflicting views, and ever increasing doubts, on its reliability. Some studies have found high inter-rater reliability, sometimes as high as .9 (Follman & Anderson 1967 cited in Perkins 1983; Moslemi 1975; Flahive & Snow 1980; Kaczmarek 1980; Mullen 1980; Jacobs, Zinkgraf, Wormuth, Hartfiel & Hughey 1981; Homburg

1984; Stansfield 1986; Stansfield & Ross 1988). Other studies report poor or non-existent correlations between raters (Remondino 1959 cited in Perkins 1983; Diederich, French & Carlton 1961; Mathews 1990; Raimes 1990; Phillip 1994; Wu 1995; Polio 1997).

In addition to the obvious cases where there is a lack of agreement between raters, Connor & Linton (1995) make the point that agreement between raters may be superficial and may mask important differences. They report that although US and Japanese EFL instructors appeared to agree on essay scores, the agreement on overall score masked disagreement on what constituted 'good' and 'poor' writing. Since lack of agreement between raters is frequently reported, it seems important to discover how and why they disagree. There now follows a discussion of the following rating issues:
i) differences in focus on form versus content, ii) NS and NNS ratings, iii) evaluations of experienced and inexperienced raters, and iv) rater training.

2.3.2.1 Form versus content

A prime concern in the evaluation of writing has been how raters manage the relative weighting of form versus content. There is some evidence that ESL teachers are less concerned with content than subject teachers. Comparative evaluations of accuracy versus coherence of writing were investigated by Lukmani (1993a) in evaluations of examination essays in Economics, Logic and Zoology at the University of Bombay. She found that an improvement in linguistic accuracy was the key factor that caused the ratings of the ESL teachers to rise, while a joint improvement in coherence and linguistic accuracy was what caused the subject teachers' ratings to significantly rise. The finding that ESL teachers valued content less than form was supported by other studies (Bridgeman & Carlson 1991 in Hamp-Lyons 1991d; Song & Caruso 1996).

In the light of these findings, it is interesting to note that ELTS readers were told to apply criteria that were not discipline based and to ignore inaccurate content (Hamp-Lyons 1991d). In the same study it was noted that readers were concerned, not so much with quality of content, as with a convincing discourse structure. The four readers had masters degrees in TESOL and were evaluating 23 ELTS essays. This concern with form rather than with content is frequently signalled to ESL students, both

by the kind of feedback they receive for their essays, as well as information given to them in advance that their essays will be evaluated according to quality of form rather than content. For example the students Homburg (1984) investigated on the University of Michigan rating scale were told that mechanics and comprehensibility would be valued, not content.

In contrast to the studies cited above, Santos (1988) found that the 178 professors' evaluations of two ESL students' writing seemed to focus more on errors than on content. What is obviously important is not whether content should be valued more highly than accuracy or vice versa, but how the relationship works between the two. Hamp-Lyons (1991e) makes the point that when the writing fails to engage the reader, it is at this point that language problems are noticed. Although there is evidence that writing teachers tend to be suspicious of, or confused by, technical vocabulary and do not respond well to discipline based writing, there are some indications that this changes slightly when the writers are more skilful (Hamp-Lyons 1991d). It is clear that the level of the writing will make a significant difference to the aspects of writing that receive most attention.

2.3.2.2 NS versus NNS raters

Differences between native speakers' and non-native speakers' evaluations of English have been proposed by Kaplan (1966) and others. Two kinds of difference are hypothesised between the groups. One is the way a non-native speaker group may take into account their own culture-specific pattern of writing (Basham & Kwachka 1991). Kaplan (1967: 15) suggested that '..rhetoric...is as much a culturally coded phenomenon as the syntactic units themselves are'. The other is NS and NNS assessors' ability to assess linguistic accuracy. The concern over NNS ability to assess (and teach) linguistic accuracy is a matter for political concern since many university departments seeking ESL teachers have a policy of employing either solely or preferentially NS teachers. Research findings seem to present a confusing picture on this issue. There seem to be differences of opinion both on whether or not there are clear differences between NS and NNS groups of raters, as well as disagreements as to what the differences are. When differences are found, they often vary from one group of subjects to another.

Substantial differences between NS and NNS ratings were found by Hinkel (1994), whereas Hughes and Lascaratou (1982 cited in Davies 1983) found no differences between them. Phillip (1994) compared NS and NNS raters and found individual rather than group differences, but found that the measure of greatest agreement among all raters whether NS or NNS was in the area of linguistic accuracy. In contrast, several studies found little agreement between NS and NNS raters on accuracy, but found instead that NNS raters were generally more severe in their marking (James 1977; Davies 1983; Santos 1988; McCretton & Rider 1993). Other studies reported that NS and NNS raters were severe in different areas. Some studies found that NNS teachers were found to rank mechanical errors as the most important criterion in evaluating student writing (Applebee 1981 cited in Astika 1993; Zamel 1985; Green & Hecht 1986). On the other hand, Kobayashi (1992) found that native speakers of English were more strict about grammaticality, but more positive in their evaluations of clarity of meaning than their Japanese counterparts. James (1977) supported Richards' (1971 cited in James 1977) finding that NSs care more about verb morphology than they do about lexical errors, but Khalil (1985) found that the native speaker evaluators he studied penalised semantically deviant utterances more harshly than grammatically deviant ones. Green & Hecht (1986) commented that NS and NNS raters had difficulty in agreeing both on what constituted an error, as well as on the gravity of various errors. Maybe the most sensible view is McCretton & Rider's (1993) conclusion that there is no universal error hierarchy as far as evaluation is concerned. They thought that any error hierarchy was influenced most of all by the rater's education. Another factor to take into account is culture.

Culture may cause second language writers to put different values on knowledge and how it is organised and an assessor from a different culture may not understand those values (Clyne 1981; Ballard & Clanchy 1991). Carrell (1983) emphasised that the readability of text depended on the rater's general knowledge of the world and the extent to which that knowledge was shared. Chichara, Sakurai and Oller (1989) concluded that culturally determined expediencies were more important than syntactic complexity and Basham & Kwachka commented:

> 'We have found that the cumulative effect [of cultural views] on the tone of the writing dominates the assessment of the extent that an instructor or assessor may actually be surprised, on close reading, to find a fairly well organised and coherent statement.' (1991: 43)

Hamp-Lyons (1991a) reported that raters evaluating the ELTS test were influenced by cultural background, but Khalil (1985) found that NS raters' ability to interpret the texts of Arab EFL writers was not influenced by differences in cultural background.

One aspect of culture conditioning is the effect it can have on the type of English that is used. Where English is used extensively, a local standard English may be produced. This is the case in Papua New Guinea, where the thesis research was conducted, and so the issue of whether language varieties of English, other than the standard British, American or Australian should be accepted, arises. Hyland (1990), writing on communicative competence in Papua New Guinea schools, argued that PNG-English was here to stay and should be used and accepted as a legitimate English language variety. The issue of varieties of English is complex and in Papua New Guinea there is presently no consensus among its educators as to whether PNG English should be formally accepted or not. This obviously affects text evaluation, as it must do in similar situations elsewhere.

2.3.2.3 Experienced versus inexperienced raters

Another reason for differences in evaluation may be the differences in expertise between experienced raters and those who have only little experience, although there are conflicting findings. For example, Cumming (1990) found that there were significant differences in the way experienced versus inexperienced raters evaluated content and rhetorical organisation, but not in the way they evaluated language use. Schoonen, Vergeer and Eiting (1997) found the opposite. They compared expert versus lay readers and found that the experts were much better at rating language usage, but that both lay and expert readers agreed on evaluation of content. Santos (1988) and Vann, Meyer and Lorenz (1984) investigated rater harshness and found that age and experience caused older raters to assess more leniently than younger, less experienced colleagues. Shohamy, Gordon and Kraemer (1992) found that rater age and background did not matter but that rater training and experience did.

2.3.2.4 Rater training

Rater training is advocated by many researchers in an effort to maximise reliability (Hughes 1989; Shohamy, Gordon & Kraemer 1992; McIntyre 1993; Weigle 1994; Lumley & McNamara 1995),

although all agree that variability still remains. Hamp-Lyons (1991f) reported that inter-rater reliability on the TWE where both ESL trained and English subject teachers evaluated the scripts was 0.67 - 0.72, while MELAB's trained raters achieved an inter-rater reliability of 0.90. The problem with rater training is that the differences in evaluation discussed above show that raters seem to have personalised sets of criteria with which to judge pieces of writing.

Error, for example, as well as other facets of writing, is evaluated differently by different raters. Vann, Meyer and Lorenz (1991) drew attention to the fact that response to errors is more complex than previously thought and that evaluation was not just a case of the quantity or quality of error. They emphasised that evaluation depended on rater's often highly personal criteria. It is these sets of personal criteria for judging writing which become invalidated by the attempt to impose scoring scales with preset fixed criteria during rater training (Huot 1990). Charney (1984) pointed out that rater training sessions use peer pressure, monitoring and rating speed to make sure raters use the 'right' criteria. As Barritt, Stock & Clark (1986) emphasise, reliability does not equal validity. They recommend that we should acknowledge that well informed judgements can be inconsistent and that raters should not be trained to ignore their own experience.

Another problem with rater training is that there can still be substantial variation in rater harshness (Charney 1984, Lumley & McNamara 1995) or the training can even appear to have had no effect at all (Carlisle & McKenna 1991; Vaughan 1991). In addition, Lumley & McNamara (1995) noted that rater characteristics were not always consistent over time. They recommended that multi-faceted Rasch measurement be used, where possible, or at least multiple ratings, to compensate for the lack of effectiveness of rater training. Kroll (1998) notes that there is a great need for further research into why raters differ.

Some increase in reliability has been reported as a result of rater training (Stansfield 1986; McIntyre 1993) but as noted above, reliability does not equal validity. Such reliability has presumably been achieved by the domination of one set of criteria over others. The TWE, for example, seems to achieve its reliability partly through the testing of raters to see if they can mark according to previously set

criteria, and if they cannot, they are not employed (Stansfield 1986). Doubts as to the validity of the assessment when raters are trained to conform to predetermined values are expressed by many researchers (Charney 1984; Barritt, Stock & Clark 1986; Huot 1990; Henning 1991; Horowitz 1991). Wiegle (1994) comments that although rater training helped clarify scoring criteria and provided a reference group of other raters with whom they could compare themselves, it was precisely the existence of the group that might adversely affect the scoring. This was both because of peer pressure to grade similarly and because dominant personalities usually carried more weight than other considerations in agreeing interpretations of criteria. The question is: who knows best? And if we are not sure, is it ethical to impose arbitrary scoring scales?

2.3.3 Task variability

Task variability is the second major source of concern in the achievement of reliability of measurement. Hamp-Lyons (1991e) draws attention to the need to specify audience, to control for topic familiarity and discourse type, and she points out that there is no agreement on whether or not it is better to provide a choice of task. Providing a choice would enable writers to choose a subject that was familiar to them, but could create reliability problems because the topics might elicit different levels of performance. Tedick (1990) found that giving ESL students a field topic rather than a general one discriminated better, but Reid (1990) commented that the responses to different task demands, although measurably different, were not always easy to either predict or to quantify. Kroll and Reid (1994) emphasise how important it is to trial the test prompt.

Prompt difficulty depends on audience specification, topic familiarity, genre and the complexity of the prompt's syntax (Hamp-Lyons 1991e), but the challenge in a task, the interest it generates in the student and the motivation to write can override other areas of difficulty. Hamp-Lyons makes the point that: '...all writing...is creative and personal as well as communicative...' (1991a: 52) and so the kind of topic that is given is a crucial factor in whether or not the student is motivated to write. Students need to make a topic their own in order for it to be motivating, and ESL students need to feel that it is possible to write about. There is evidence, too, that prompts do not work in the same way for L1 as for L2 students (Hamp-Lyons & Prochnow 1990 cited in Hamp-Lyons 1991f).

Horowitz (1989) considered that essay examination prompts constitute in some ways a separate genre, since they have a shared set of communicative purposes. He drew attention to the hidden rules of such tasks, for example, that the student should write about what was learned and not what was known before, that the student should not express an opinion unless asked, and that the student should pretend that the examiner did not know the information that was being given in the answer. The artificial readership provided by the educational setting changes the writing purpose in that the criteria for the success of the piece of writing are imposed from without by the prompt setter and evaluator (Kroll and Reid 1994; Reid and Kroll 1995).

The level of cognitive difficulty implied by particular tasks must obviously have some effect (Hamp-Lyons 1991c). Zhang (1987) found that the level of cognitive complexity of the question influenced the syntactic complexity of the writing and the number of words produced, but did not cause a significant difference in the level of accuracy. To add confusion to the effect of the level of cognitive difficulty contained in the prompt, Alderson and Lukmani (1989) found that nine expert judges could not agree, in more than half the cases, on which level of complexity a question was testing. Hamp-Lyons and Prochnow (1994) found, too, that assumptions about task difficulty were not always correct and were sometimes the reverse of what was predicted. Other research contradicted these findings, however. For example, Brutten, Perkins & Upshur (1991 cited in Lumley 1993) found substantial agreement between four raters and Lumley (1993) found that five experienced ESL teachers achieved substantial agreement on question levels.

It is known, too, that students can perform differently on different types of discourse. For example, Crowhurst & Piche (1979) found that there was significant variation in syntactic complexity associated with mode of discourse. It follows that it is dangerous to assume that writing types are equivalent, as happened with the TWE test where the 'compare and contrast' task was equated with the 'describe and interpret' task, but where the correlations showed them to be not equivalent (Hamp-Lyons 1991c). Kroll (1998) warns that 'good' writing varies according to text type and that the kind of writing that is successful for one type, may not be for another. On the other hand, Hamp-Lyons (1991e) reported that

some studies had found correlations as strong across topic types as within topic types. It seems that we should proceed with caution.

The writer's response to the topic is affected by experience, physical and emotional state at the time of writing, and, some think (e.g. Carrell and Munroe 1993; Horning 1993), by personality. Carrell and Munroe (1993) claim that 'feelers' write more lexically diverse pieces, while 'thinkers' produce writing that is more easily assessed in a consistent fashion by raters. The writer's response is difficult both to assess and to control. A few researchers have taken into account the variable states of writers and have suggested that single drafts are not adequate evidence on which to base a writing assessment (Purves 1992, Lumley & McNamara 1995). For most tests, however, whether school-based, or large-scale, the practical difficulties involved in alleviating this source of variability are too great to consider seriously. What is clear, however, is that writers do have personal responses to topic types and that these can have an effect on performance.

2.4 Objective indicators of writing development

Objective measurement is not usually used for the same purpose as subjective evaluation. Subjective evaluation tends to be used for large scale testing to give holistic scores that are suitable for placement purposes, while objective measurement tends to be used either in conjunction with subjective evaluation or by itself in order to identify features that contribute to text quality and which therefore seem to indicate writing development.

There are difficulties with objective measurement. Firstly, objective measures that count different types of features have a problem in that there is an assumption that quantity matters, when it may not. Secondly, objective measurements are sometimes tied to subjective evaluations in studies that assess which objective features discriminate between levels of proficiency. The levels of proficiency have been identified by subjective ratings, which means that the problems of subjective evaluation have to be taken into account. Thirdly, objective measurements are not as objective as they might seem and are not comprehensive, i.e. they do not measure the whole text. Decisions on items and answers, and on methods of analysis are made beforehand. In other words, decisions as to what is important and what is

not in the evaluation of writing are made subjectively by the test constructors (Hamp-Lyons 1991c). Validity, therefore, seems to be compromised by objective measurement, although reliability can be increased (Polio 1997).

The mapping of growth in writing through a description of objective features that seem to be associated with increased writing competence has identified several likely candidates. Measures of syntactic maturity, such as Hunt's t-unit (Hunt 1983), used either alone or combined with measures of accuracy (Larsen-Freeman & Strom 1977; Larsen-Freeman 1978; Perkins 1980), measures of vocabulary, and investigations into the number and kinds of cohesive tie (Evola, Mamer & Lentz 1980, Witte & Faigley 1981) have all sought to identify indications of writing development. For persuasive writing, attempts have been made to quantify the levels of abstraction students incorporate into their writing in order to provide indices of writing competence development. More recently, several researchers have come to the conclusion that one of the best measures to significantly indicate the development of students' writing competence is simply the amount of fluency they demonstrate, measured by the number of words they produce (Carlisle & McKenna 1991; Ferris 1994; Kamimura 1997).

Findings on objective indicators of writing development will be reviewed and discussed. Studies are often difficult to compare because of the variability of subjects' backgrounds and levels of language proficiency. They are also difficult to compare because descriptions of methodology sometimes fail to give information on the precise kind of writing used for the study, or results may omit findings on inter-rater reliability where subjective evaluation was used in conjunction with the objective measure. (See Appendix A for summaries of 33 of the studies that are discussed in this section.)

2.4.1 Measures of syntax

The most commonly used measure of syntax is the t-unit, which is widely used as a base measure of syntactic development in writing. Hunt (1965 cited in Hunt 1983) developed the t-unit as a measure for first language writing. He found (1970 cited in Hunt 1983) that the average number of t-units in student writing gradually decreased as students progressed through school and became better at

punctuation and as they attempted harder genres like persuasive writing. A concurrent development was that the length of t-unit gradually increased, as students wrote more formal English requiring increased nominalisation and subordination. This phenomenon was found to hold across languages as experiments were carried out in languages such as Fijian, Korean, Indonesian (Loban 1976 cited in Watson 1983; Hunt 1983), although Witte (1983) emphasised that t-unit length stabilised only after repeated practice of writing skills.

There are, however, problems regarding the use of the t-unit as a yardstick. The first concerns a criticism that writers of international standing, for example Hemingway and Faulkner, were capable of scoring very low as well as very high on the syntactic maturity scale (Britton 1975; Hunt 1983). The second concerns variability influenced by discourse type, audience and topic. Many studies showed that changes were attributable to discourse type as well as to age, or maturity as a writer (Fosen 1969 cited in Watson 1983; Crowhurst and Piche 1979; Rubin and Piche 1979; Freedman & Pringle 1980; Miller 1980 cited in Watson 1983; Witte 1983). The t-unit was adopted as a useful measure for second language writing research and the average length of the t-unit was found initially (Larsen-Freeman & Strom 1977) to be one of the best measures to use in constructing an ESL index of development. Further research to develop the index, however, revealed that an error-free t-unit was a better measure (Larsen-Freeman 1978). Perkins (1980) investigated the relationship of objective measures to holistic evaluations of ESL writing and concluded that it was only error-free measures that discriminated among levels of proficiency. There is no consensus on whether error-free measures are superior or not. Some studies of ESL writing found error-free measures to work better (Larsen-Freeman 1978; Perkins 1980; Ishikawa 1995) while others found measures that included error to be better indicators (Flahive & Snow 1980; Intaraprawat & Steffensen 1995). Homburg (1984), in an analysis of ESL writing for university placement tests, found both t-unit length that included error and the number of error-free t-units per composition to be significant discriminators.

Both first and second language writing research has taken to heart the t-unit research message that writing development tends to be associated with increased complexity of sentence structure, and has developed other measures of writing complexity to see if they are more powerful. Following Hunt's

work, Christensen (1968) analysed the L1 writing of both college students and professional writers and came to the conclusion that free modifiers, and in particular, final free modifiers were the mark of skilled writers rather than simply t-unit length. Christensen defines free modifiers as follows:

> 'In the initial position all words and constructions that stand before the noun phrase that is the subject are free modifiers regardless of punctuation; position alone marks them as free... Every medial or final word or construction that is set off is a free modifier.' (1968: 578)

He also gives examples with the free modifiers in italics:

> '....*When the Times wanted to transfer him to Bonn, a bigger story and a bigger bureau,*[2] he went reluctantly, *leaving what he had come to call "my people."*[3] We shared, *I think,*[4] the same feeling for being a reporter there, *of watching and in a way being involved in the simple yet moving business of the daily struggle of these people with the state.*[5]' (1968: 578)

Inspired by Christensen, Wolk (1970) also analysed the L1 writing of college students and professional writers to see how syntax changed with skill level, and found similarly that final free modifiers were a more powerful indicator than average t-unit length alone. Nold & Freedman (1977) confirmed this finding in a longitudinal study of freshman writing, where modifiers and especially final free modifiers were found to indicate writing development, although in their study t-unit length was not found to be an indicator of development.

Second language writing research seems to have been less interested in free modifiers as measures of development, probably because substantial amounts of modification are associated with an advanced level of writing skill that is not its main concern. The ratio of number of clauses to t-unit or sentence, however, has been used to measure second language writing development. The clause/t-unit ratio was found be one of the best indicators of level of ESL writing development by Flahive and Snow (1980), a finding supported by Zhang (1987) in her investigation of two levels of ESL intermediate writing. Homburg (1984) used levels 5, 6 and 7 of the Michigan Test, which he considered to be adequate for university entry level, to investigate which features of text discriminated between levels of ESL

---

[2]initial free modifier
[3]final free modifier
[4]medial free modifer
[5]final free modifier

writing. He found that the number of dependent clauses was a significant discriminating feature between all three levels.

The t-unit and error-free t-unit, however, have remained the most popular base measures. They can apparently give us a rough indication of writing maturity, although the effect of discourse type, audience and topic, as discussed above, as well as other variables, are recognised to influence the results (Watson 1983). Many offer support for the view that 'syntactic maturity', i.e. as captured by t-unit measures, is related to quality, both in L1 writing (Wolk 1970; Rubin & Piche 1979; Witte & Faigley 1981; Hunt 1983) and in ESL writing (Larsen-Freeman & Strom 1977; Larsen-Freeman 1978; Flahive & Snow 1980; Homburg 1984; Carlisle & McKenna 1991; Casanave 1994; Intaraprawat & Steffensen 1995).

2.4.2 Measures of accuracy

Measures of accuracy are felt intuitively by both teachers and researchers to be related to text quality (Larsen-Freeman 1978). This makes sense in view of the fact that it is recognised that the development of writing competence necessitates a gradual automatisation of lower level skills such as handwriting, spelling, punctuation and common syntactic forms (Cummins 1983). This gradually lessens the load on STM, leaving more processing space to devote to overall writing goals. The precise relationship between lack of error and quality, however, is difficult to quantify. One reason for this is that teachers and evaluators perceive different errors as intrusive (Connors & Lunsford 1988). Another reason is that frequency of error is not always related to seriousness of error. A third reason is that other features of the text seem to override considerations of language error in certain circumstances.

Measures of accuracy involve counting errors. Identification and classification of error, however, remain problematic (Homburg 1984). James (1974) comments that all errors do not have the same weight, and that frequency is not necessarily relevant to gravity. Polio (1997) reports five reasons for disagreements over error classification:

1. legibility
2. questionable prescriptive language rules

3. questionable native-like usage
4. intended meaning unclear
5. mistakes on the part of the raters

Rifkin & Roberts (1995), in a review of error gravity research design, recommended that error gravity research should be reconceptualised and earlier studies reassessed because the process of error evaluation has to be shaped by extra-linguistic factors.

Research into second language writing development puts an understandably greater emphasis on accuracy than does first language research, since second language writers are likely to make more language errors than L1 writers and errors are easily visible items that seem capable of correction. Both fields, however, have been concerned with the role accuracy plays in the contribution to text quality and writing development. Measures of accuracy can be problematic in terms of identification and classification. Despite emphasising this point, Polio (1997), who reviewed 16 studies of linguistic accuracy as well as reporting her own, found that measures of error could be reliable. Problems arise, however, not so much from doubts about reliability, as from the difficulty of comparing studies that have identified and classified errors in different ways.

Although there is general agreement that as competence develops, writing tends to contain fewer language errors, it seems that the straightforward counting of number of errors may not discriminate clearly between levels of proficiency. Tarone, Downing, Cohen, Gillette and Murie (1993) found that there was no significant difference in the levels of accuracy of ESL 8th, 10th and 12th graders. Two other studies of ESL writing also found that overall accuracy was not a predictor of quality or of writing development (Flahive & Snow 1980; Zhang 1987). In a comparison of ESL, EFL and L1 writing, Carlisle and McKenna (1991) found that the measure of overall accuracy was not an indicator of quality for either NS or NNS writing and despite the fact that the NNS writing contained significantly more errors than the NS writing. They found that the level of accuracy did not seem to influence raters.

Other researchers, however, have found otherwise and claim that accuracy does contribute significantly to text quality. For example, Sweedler-Brown (1993) also compared L1 and ESL writing and found that corrected ESL essays were rated higher than the originals which contained errors. She found, unlike Carlisle and McKenna (1991) cited above, that raters did pay attention to accuracy and that accuracy was a significant indicator of writing development and quality. For L1 writing at first-year college level, Witte and Faigley (1981) found that overall accuracy was a predictor of text quality. For ESL writing, Homburg (1984) found similarly that accuracy was a predictor of text quality. From levels 5 - 7 on the Michigan Test the average number of errors dropped at each successively higher level of proficiency.

The level of ESL writing development seems to influence both the number and type of errors that students make in their essays. Homburg had earlier (1981) investigated the number of errors made at different levels of the Michigan Scale - which could be equated with levels of writing competence development. The errors were divided into three categories, according to how serious they were considered to be. First degree errors, the most serious type, increased dramatically between levels 5 and 6 indicating a jump in risk taking, but started decreasing at level 7. The number of second and third degree errors, the less serious kinds, remained stable between levels 5 and 6 and decreased to almost zero by level 7. The findings support the common sense view that the number of errors seems to change according to the level of development, and may initially rise indicating perhaps a willingness to take risks, before dropping again. Casanave (1994) in a longitudinal study of journal writing points out that almost half of her intermediate level ESL subjects wrote less accurately as the academic year progressed. Larsen-Freeman and Strom (1977) pointed out that good writing often had more errors than poor writing and this comment may refer to writers at an intermediate level of proficiency, whose risk taking is increasing and whose essays are becoming interesting and well-formed in other ways. There remains, however, a general consensus that error is a significant indicator of ESL writing proficiency (Perkins 1980, 1983).

Specific types of error have been claimed to indicate writing development or text quality. Scott & Tucker (1974) investigated the narrative writing development of ESL low intermediate students. They

found that as competence developed, significantly fewer errors were made with finite verbs and prepositions, but that subject-verb agreement problems and errors with articles did not change much. Vann, Meyer & Lorenz (1984) studied faculty evaluation of ESL errors and came to the conclusion that there was a hierarchy of errors, where word order, tense and relative clause errors were considered most serious, while errors with article usage and spelling errors were considered tolerable. Vann, Meyer and Lorenz later (1991) refined and extended their study and produced similar findings on the issue of the relative importance of various types of error, i.e. verb forms most serious, article errors less serious and spelling errors least serious, although they emphasise that response to writing is complex and depends on more than the quality and quantity of error.

In contrast to the findings of Vann, Meyer and Lorenz (1991), spelling errors have been found in some studies of the development of L1 writing skills spanning a range from children to young adults, to strongly affect ratings of text quality and writing development. For example, Grobe (1981), in an investigation of writing development of native speaker writers at grades 5, 8 and 11, found lack of spelling errors to be an indication of writing development in both narrative and expository writing types. In another study of L1 writing to investigate the relationship of holistic scores to objective features, Charney (1984) found that spelling errors were a significant indicator of quality. It is understandable not only that spelling would appear more important at lower levels of proficiency, but also that it would have more prominence in L1 writing where other types of error presumably occur less.

The type of task and especially the audience requirement, as well as the level of development, are known to have an effect on the number of errors. Persuasive writing is recognised to be more difficult than narrative or explanatory writing, for example, so that students make more mistakes as they grapple with the more complex sentence structure required (Watson 1983). The degree of intimacy of the intended audience, even within a single writing type, affects the level of difficulty, too (Rubin & Piche 1979).

Unfortunately, many ESL studies use accuracy as a yardstick where it is simply assumed that accuracy is a crucial feature of quality and where conclusions based upon this assumption are drawn. Robb, Ross and Shortreed (1986), for example, investigated the effects of various types of feedback on student writing by counting the number of error-free t-units in their subjects' in-class narratives. Ishikawa (1995) compared two methods of eliciting the production of narrative using the same method. Carlisle (1989 in Polio 1997) investigated the comparative effects of a bilingual versus a submersion programme by counting the number of errors in student scripts.

It is clear from Polio's (1997) findings, as well as from my own experience, that the writing type, conditions for writing and methodology are often not fully described in descriptions of studies, so that it is difficult to make comparisons between them. Polio also noted the difficulty of finding research on this subject, since academic databases do not classify measures of linguistic accuracy as a search topic. Such difficulties make the gathering, evaluation and comparison of findings on the contribution of accuracy, problematic. The same difficulties apply to studies investigating the contribution of other text features to the development of writing competence, for example, the contribution of vocabulary and content which will be considered next.

2.4.3 Vocabulary and content

Some studies lend support to the view that content should be valued more highly than form with findings that semantic errors, such as poor word choice or illogical statements, can interfere more with intelligibility than grammatical errors (Khalil 1985; Santos 1988). Some researchers have found measures of vocabulary to be one of the best indicators of writing development. In a study of L1 writing, Charney (1984) found that unusual words could positively affect ratings, and other studies supported the view that vocabulary was a significant measure of quality (Nold & Freeman 1977; Grobe 1981). In ESL writing Mullen (1980) found, in a study of 117 samples using five scales, that vocabulary was the best predictor of quality. Astika (1993), in a study of 210 samples of L2 writing using four topics, found that vocabulary was the best predictor. In an analysis of ESL placement tests Ferris (1994) found vocabulary and fluency to be the only measures that discriminated significantly between levels of proficiency.

Despite such findings and the repeated emphasis by ESL researchers on how important it is to value content rather than merely concentrating on form, there seem to be very few studies in lexical development. Laufer (1994) studied the lexical progress of advanced ESL writers over one academic year. She found some progress in lexical richness, but not in lexical variation. It is worth remembering that different genres have differing requirements for lexical variation, with academic writing needing less variation than narrative writing. Given this observation it is not surprising that little increase in lexical variation occurred in the academic prose of advanced writers as they developed their competence. Engber (1995), however, investigated intermediate to advanced ESL students in a cross-sectional study and found that the measure of lexical variation, both error-free and otherwise, correlated with holistic scores. It is possible that Laufer's (1994) study of a year's writing development might not have been long enough to show development in lexical variation, or that the lexical variation used by her advanced writers had already peaked and stabilised. In a later study, Laufer and Nation (1995) re-emphasised that choice of lexis significantly influenced raters' evaluations. This view was supported by Santos's (1988) finding that lexical errors were considered the most serious by the 178 professors whose evaluations of ESL essays she compared.

Another way of measuring the development of the quality of content in persuasive writing has been to investigate the levels of abstraction students use. Freedman and Pringle (1980), in their account of indices of L1 writing growth in the college years, concluded that the level of abstraction seemed to be an indication of growth. It has been found to be an important indicator for ESL writing, too. Connor (1991) found, in an investigation of persuasive essays written for the Test in Written English, that 59% of the variance on holistic scores of 22 ESL students could be accounted for by the incidence of the 'warrant' category (bridge between data and claim) on the Toulmin scale of reasoning. The scale was developed to measure the logical constituents of a persuasive essay. The importance of text content, however, cannot be considered without taking into account the links between ideas that make the topic coherent.

2.4.4 Cohesion and coherence

Measures of cohesion have been proposed as significant features of writing growth. This is based on a growing awareness that the relative merits of fluency, accuracy and the complexity of syntax are subordinate to the concerns of cohesion and coherence. Problems of cohesion and coherence are particularly visible when immature writers attempt academic or non-narrative writing, where the text structure is more complex. Unfortunately, it seems that although we intuitively recognise the importance of coherent, cohesive text, it is hard to measure and confirm objectively. Studies that have focussed exclusively on measures of cohesion and coherence have provided insights into what makes texts cohesive and coherent and into problems that writers have with their text organisation, but studies that have correlated markers of cohesion and coherence with holistic ratings, with the exception of a study by Intaraprawat and Steffensen (1995), have generally found that cohesion and coherence markers are not clear indicators of writing development (Evola, Mamer and Lentz 1980; Freedman & Pringle 1980; Mullen 1980; Witte & Faigley 1981; Homburg 1984).

Bamberg (1983) investigated the question of what makes a coherent text by studying the descriptive writing of high school L1 writers. She concluded that almost any text feature can contribute to coherence, but that it is useful to differentiate between 'global coherence', an overall consistency of goal and text, and 'local coherence' which she described as local problems within the text. Bamberg stated that problems that affected global coherence were far more important than local problems, such as mechanical errors. She also made the point that cohesive ties were not sufficient to create coherent text, which Carrell had discussed earlier in her article 'Cohesion is not coherence' (1982). In this article Carrell made the points that cohesion is not the cause of coherence, but its effect, and that coherence is achieved through a shared context between writer and reader. The identification of this area of difficulty for L2 writing was supported by empirical findings, which demonstrated that L2 writers experienced difficulty in producing coherent text (Zamel 1983b; McDevitt 1989; Johnson 1992).

Bacha and Hanania (1980) conducted an experiment with 300 L2 learners in the American University of Beirut and found that specific teaching of linking words produced a 7.3% improvement in the

writing and they found that punctuation improved more than any other single aspect of the students' writing. Their finding that the increased ability of students to use correct and effective punctuation seems to be associated with a development of writing competence in academic writing is not surprising. Presumably, since the typical form of written language relies on subordination, i.e. on pieces of text referring to, relating to, and being subsumed under, other pieces of text, then an effective use of commas would contribute heavily to the clarity of the language. What *is* possibly surprising is that punctuation, as far as we know, was not specifically taught or focussed upon. As far as the hierarchy of linking words is concerned, Bacha and Hanania (1980) found the most needed transitional words were also the easiest, while the least needed words i.e. those most infrequently attempted, were the most difficult. They found the following apparent order of acquisition:

1. linking words showing result, reason and addition - most needed and easiest;
2. linking words showing comparison and contrast - less needed and less easy;
3. linking words used to show clarification - least needed and most difficult.

Studies which have attempted to link cohesion and coherence markers to holistic ratings, however, have generally found that the measures were not significant indicators of writing development. In a study of the writing development of native speaker college students Freedman & Pringle (1980) found that there was a significant difference in rhetorical skills between levels of proficiency but that these broke down as harder genres were attempted. Witte and Faigley (1981) concluded that for L1 writers, measures of cohesion were not significantly correlated with writing quality, but acknowledged that cohesive links were, nevertheless, important. A study on the writing development of ESL 8th, 10th and 12th graders using narrative writing samples (Tarone, Downing, Cohen, Gillette & Murie 1993) found no significant differences in measures of coherence between the grade levels, although there was a difference related to the number of years the subjects had spent in the U.S. In a study of ESL writing development (Homburg 1984) where writers were approaching readiness for university study, there was an uneven development of rhetorical skills. Evola, Mamer and Lentz (1980) analysed the essays of 94 Arabic and Farsi speakers to investigate their use of cohesive devices and found that skill in using cohesive devices was a minimal indicator of overall language proficiency. Mullen (1980) had similar

findings, when she reported that the ratings on organisation for 117 NNS scripts were the worst predictors of quality.

In contrast to the studies cited above, Intaraprawat and Steffensen (1995) investigated the use of metadiscourse features (e.g. connectives, code glosses, illocutionary markers, hedges etc.) as indicators of quality in the persuasive essays of ESL university students and found that the good essays contained twice the density of metadiscourse features. For example, four of the six good essays used every type of metadiscourse, but none of the poor essays did. Most studies, however, of both L1 and ESL writing, as discussed above, found that objective measures of cohesion did not appear to discriminate well between levels of writing proficiency. Such findings may be partly explained by Carrell's (1982) view that coherence depends on more than the number of cohesive ties and Nystrand's (1982b) emphasis that coherence depends not only on cohesive ties in the text but on the nature of the textual space shared by writer and reader. Nystrand's comment may explain why some studies have found that rater agreement on text organisation has been more difficult to achieve than agreement on other text features (for ESL writing: Mullen 1980; Astika 1993; Phillip 1994; for L1 & ESL writing: Sweedler-Brown 1993).

2.4.5 Fluency

Perhaps surprisingly, several researchers have found that the fluency measure, i.e. the number of words produced overall, is one of the best indicators of writing development and text quality both for L1 writing (Nold & Freeman 1977; Grobe 1981; Witte & Faigley 1981; Carlisle & McKenna 1991) as well as for ESL/EFL writing (Winfield and Barnes-Felfeli 1982; Zhang 1987; Carlisle & McKenna 1991; Ferris 1994; Hirose & Sasaki 1994; Intaraprawat & Steffensen 1995; Kamimura 1997). Skehan (1996) offers the explanation that fluency is important because it takes time to express interesting ideas.

Most studies have found that length is a significant predictor of holistic scores, i.e. of quality in the writing, although not all agree. Some studies have found fluency not to discriminate significantly among holistic evaluations of ESL writing (Evola, Mamer & Lentz 1980; Perkins 1980; Connor 1991).

Homburg (1981), in his study of three levels (5, 6 and 7) on the Michigan Scale found that although fluency increased between levels 5 and 6, it stabilised at level 7. Three points need to be emphasised. Briefly these are that fluency stabilises at some point, that it may vary according to task type, and that long is not necessarily beautiful.

2.4.6 Other indicators

The text features so far discussed in this section have been features such as accuracy, fluency and cohesive links that contribute mainly to the readers' ease of text processing. They are the kind of features where quantitative measurement could be attempted. They attempt to identify a minimum acceptable writing proficiency in general terms and arise from a teacher focus on what goes wrong with student writing. This should not blind us to the fact that even in low-level student writing there are other factors that contribute to quality, which interact with concerns of linguistic accuracy and text cohesion and which may influence readers strongly at varying levels of consciousness. The features of text quality described below are not easy to measure but make a significant contribution to writing quality in the sense of their specially effective ways of conveying meaning and making text memorable.

Novelty can contribute to the quality of writing. Barritt, Stock and Clark (1986) commented that when the text was surprising, raters often did not agree, and would start discussing the writer rather than the text. There is supporting evidence that unusual words have a positive effect on ratings (Nold & Freedman 1977; Grobe 1981; Charney 1984). Something novel and unexpected in a text arouses interest and helps make the text memorable.

The use of simile and metaphor can contribute to text quality. They make text memorable partly because of the images they provide and partly because the text's meaning has to be reached in two steps rather than one. Metaphor is mentioned by Berg as '... one of a variety of interpretation operators mediating between literal meaning and speaker meaning' (1989:191). Shepherd (1994) comments on the use of simile and metaphor that connotations are often more powerful than denotative language because when the reader has to close the gap, it is often closed more powerfully.

Sound is another feature that contributes to text quality. This may be difficult to research since we, as readers, are often unaware of the rhythmic effect a piece of prose may be having on us. We may also be unaware that we are using considerations of rhythm and sound when we write. Cooper and Odell (1976) tried to investigate the importance of sound from the point of view of revisions made by professional writers and found that sound revisions occupied a relatively low place in the revision hierarchy (19%), since most revisions were made to enable ease of understanding (37%). It could be that sound and rhythm are primary when we write, so that there are fewer revisions at second stage editing, but in any case a fifth of the total number of revisions is a substantial proportion. Shepherd (1994) notes that the rhythm of a piece can contribute pauses, which can change or emphasise the meaning and that the sounds of words produce echoes in the mind. The rhythm of a piece of writing contributes to the mood and the force of the text to carry us forward. Hamp-Lyons (1991a) makes the point that readers often respond more to emotional force than to logic or reason, and emotion may be associated with sound.

2.4.7 Summary

To sum up our state of knowledge on indications of writing development, it seems that despite difficulties of comparison due to variabilities in the kinds of study conducted, there is a consensus that syntax becomes more complex as writers mature and that more complex syntax as well as more complex organisational structure is used for expository and persuasive types of writing than for narrative types. A more distant audience requirement tends to elicit more complex syntax. Fluency and accuracy can be indications of writing development, and sufficient cohesive devices and an appropriate rhetorical structure are necessary for coherence, although the definition and weightings of these measures remain problematic. It needs to be noted that these features seem to be associated primarily with text readability and that additional features which involve dual methods of imprinting meaning, such as those afforded by simile, metaphor, sound and rhythm may be significant in achieving textual memorability and excellence.

2.5 Pedagogy

The view of how writing develops is that it proceeds through an interaction between writer and reader driven by a social need to compose, as discussed above (2.1). The importance of genre becomes self-evident if such a view is accepted and the relationship between genres in terms of difficulty and varying requirements offers a logical path for the development of writing proficiency. It has influenced curriculum design in Papua New Guinea and parts of Australia , as already mentioned. The increasing awareness of the cultural diversity of L2 writing interactions and educational needs, however, is pushing writing pedagogy into new ground (Cumming 1998; Raimes 1998) and theorists are confronting multiple unanswered questions.

The composing process is recognised as recursive and messy, but there is little agreement on how to teach it. Johns (1995) criticises process writing on the basis that ESL students need help with form. She recommends genre-based teaching, while Benesch (1995) responds with a defence of process writing, saying that it depends how it is taught. Genre-based pedagogy is criticised on the grounds that it encourages an authoritarian mode of teaching in the sense that form is the focus, and text models are often imposed on students (Kress 1993 cited in Raimes 1998). To add to the confusion, the original concept of genres as relatively fixed types of writing has broken down in an increasing awareness of how difficult they are to categorise. Biber (1988 cited in Paltridge 1996) distinguishes between genre, which he says is categorised on the basis of external criteria, and text type, which is defined according to linguistic form. He claims that in Australia, genre and text type are often conflated. This happened, too, in Papua New Guinea when I was involved in helping to design the new genre-based language syllabus, since it quickly became obvious that genre alone did not work as an organising category. Traditions of ESL teaching that are longer established have also been criticised, for example, by Santos (1992) in her claim that ESL teaching lacks an ideology because it is governed primarily by pragmatic aims. More recently, error correction has been criticised as unhelpful (Truscott 1996 cited in Cumming 1998). This criticism arises partly because perception of error is now recognised to vary in valid ways (Rifkin & Roberts 1995). While the theorists argue, however, the teaching continues.

Teachers have a way of using methods of teaching that they find to work, regardless of the theories which underpin them, and to an extent, regardless of fashion. If they are prevented from doing this and forced to teach in a way they do not favour, the results are often not effective, as evidenced by Pennington, Brock and Yue's (1996) observation that both positive and negative student reactions to process writing were influenced by the teacher's attitude towards the teaching method. Cumming (1992), in a case study of three ESL teachers, found that his teachers mixed their teaching methods in order to be effective. Shi (1986 cited in Cumming 1992) claims that ESL academic writing courses have one of four orientations: 'rhetorical patterns (form), function, process, or content' (1992:31) but Cumming (1992) noted that the teachers he studied drew on all four orientations. Further evidence of mixed teaching methods is provided by Pennington, So, Hirose, Costa, Shing and Niedziefski (1997) when they note, in their cross-country study of L2 teaching in the Asia Pacific region, that teaching in Australia was the most process oriented. And yet Australia is one of the places most strongly influenced by genre-based pedagogy, which is criticised as being rigid in its adherence to form. Advances in the theory of writing are important, of course, in promoting increased understanding, and hopefully translating eventually into more effective teaching, but it is reassuring that the theory is tried and tested in the workplace of the classroom, where practical experience and common sense can combine to answer some of the questions presently being asked, or to generate new ones. It is interesting to note that it is partly the awareness of differing social and educational needs that is making clear the impossibility of a universal prescription for all. Although teachers can be informed by the experience of others, they will have to make their own decisions, based on their students' needs in context, about how to proceed effectively.

There now follows a brief review of i) the teaching of the composing process, ii) types of writing practice, and iii) the practical constraints involved in the teaching of writing.

2.5.1 The composing process

The awareness of what the process of writing entails has generated teaching that emphasises multiple drafts and rewrites. Concerns with the relative merits of fluency, accuracy and text cohesion, which are based on experience and intuition as much as on research evidence, dictate the kind of feedback

teachers give their students. The relationship of writer with self in the composing process is affected by the kind of writing task and by teacher comments that feed the relationship of writer with reader. This relationship has to be optimistic in order for students to continue their struggle down the long road of increasingly difficult writing tasks. As Hirose and Sasaki (1994) point out, confidence in writing contributes heavily to success.

2.5.1.1 Planning and first drafts

The insight that writing is seen to be a generative, recursive process has changed the traditional role that planning was thought to occupy in the writing process. There are differences in opinion on how much emphasis should be given to planning. Bereiter and Scardamalia (1983), like Cooper and Matsuhashi (1983), believe that planning not only organises thought, but generates ideas, whereas Britton (1983) considers that planning should occupy a lesser role. Global planning is emphasised by Martlew (1983) as being more important then local planning and Rose (1984) supports this view. He draws attention to the fact that inflexible plans can hinder writing, and so this emphasises the importance of having a global, overarching and, hopefully, flexible plan.

In practice there are two problems associated with planning and first drafts. The first is the need to generate ideas and the second is the need to keep the overall plan in mind throughout so that it can be changed if necessary and all sections accommodated. Starting the process of generating ideas is not always easy. There is an awareness that it is important to have something to say (Britton 1975; Bereiter & Scardamalia 1983), so that interest in the topic is crucial. A lack of interest makes it hard to start writing. Several methods of helping writers overcome the blank page have been suggested. Jacobs (1986) recommends 'quickwriting', which involves concentrating on content, ignoring form and writing without stopping. Britton (1975) and Bereiter and Scardamalia (1983) advocate a 'running jump' strategy, while Wagner (1990) recommends a variety of strategies including topic-based and goal-based planning, brainstorming and talking to potential readers.

Rose (1984) has made a significant contribution to knowledge in this area with his work on writer's block. He thinks that one of the main causes of inability to write comes from having an inflexible plan

that cannot accommodate the generative nature of the process, but from a pedagogical point of view this may be hard to deal with. Planning tends to be taught where a plan for the writing is considered useful, and if the teachers state at the outset that the plan need not be followed, students tend to get a mixed message. With planning, as with other aspects of the writing process, it is the students' own perception of usefulness that counts. If forced to make a plan when they do not see a need for one, they may make it and not follow it, or they may produce the plan afterwards. The type of writing will influence the extent to which a plan may be useful, since it is obviously necessary to include certain sections of information for a report, or to plan a logical progression of ideas with evidence to support them for a persuasive essay. On the other hand, a writer might change, during the writing process, the content and storyline of a narrative text with no detriment to its purpose or success.

2.5.1.2 Feedback

Feedback is acknowledged to be crucial to the writing process. A negative response is profoundly discouraging and can inhibit writing (Hedgcock & Lefkowitz 1994; Ferris 1995). Feedback has been handled variously but the general point applies that teachers function often as the sole audience for the writing, and that they function as evaluators (Kroll & Reid 1994; Hedgcock & Lefkowitz 1994; Reid & Kroll 1995).

The most common type of teacher feedback is provided by marking errors. Davies (1983) makes two points that teachers should bear in mind. The first is that teachers should consider those errors that interrupt not only comprehensibility but which have emotional impact, i.e. grammatical errors. Students should be alerted to these. The second is that teachers should concentrate on those errors that can be remedied, hopefully, by teaching. In a large scale study of how teachers responded to errors in first language teaching in the US in 1938 and 1939, Hodges (1939, in Connors & Lunsford 1988) reported that teachers do not mark as many errors as they think they do. In his study, they marked 43% of errors and they marked these according to seriousness of error and ease of explanation. Hedgcock and Lefkowitz (1994), in a comparative study of the preferred feedback of ESL and EFL students, found that although both groups used feedback on formal text features more than any other kind, the

ESL group valued comments on organisation and content. Ferris (1995), too, found that feedback on grammar was used more than feedback on content and organisation.

There seems to be an important difference between the kind of feedback L1 and L2 students receive. Teachers of ESL students still concentrate on accuracy and form, despite exhortations to do otherwise (Intaraprawat & Steffensen 1995). Cohen and Cavalcanti (1990) note that one of the EFL teachers in their investigation reported that she did not comment on content, because it was not assessed in the examination. This probably explains the difference in feedback patterns. ESL examinations worldwide tend to stress accuracy and subordinate content. This emphasis on accuracy causes many students to see teachers as assessors rather than helpers (Cohen & Cavalcanti 1990). In contrast, the university faculty evaluators in Janopoullos' (1992) study were tolerant of L2 writing errors. They were, in fact, more tolerant of L2 writing errors than of L1 errors, which Janopoullos did not consider to be helpful for ESL students.

One of the problems of writing is the need to engage with the audience. This is made harder in pedagogical situations because '... the student's real audience is often an impoverished one, a teacher who is considered a stickler for grammar and mechanics, not someone fundamentally interested in the ideas in the text or the development of the essay, and certainly not someone to be engaged in dialogue.' (Intaraprawat & Steffensen 1995:255). Maybe it is to alleviate this situation that the advice on written response is that it is better to concentrate on content than on form (Raimes 1979; Rouse 1979; Littlejohn & Hicks 1989; Susser 1994; Intaraprawat & Steffensen 1995). Fathman and Whalley (1990), however, investigated the relative merits of content versus form feedback and found that feedback on content improved content less than grammar feedback improved accuracy, although they found that both types helped to some extent. Kepner (1991 in Polio 1997) compared the effect of message-related feedback to error correction feedback on second year university students' journal writing and found no significant difference between the two types. Brewer (1988) found that feedback on accuracy was not helpful at all. In fact teacher feedback was widely reported to be unsatisfactory (Cohen and Cavalcanti 1988). L1 writing projects criticised teachers' feedback as being either too general (Sommers 1982 cited in Leki 1990) or for being too specific and focussing too heavily on

surface features (Searle & Dillon 1980 cited in Leki 1990). Hillocks (1986 cited in Leki 1990) reviews dozens of research findings and concludes that teacher comment had little impact on student writing. Leki (1990) reported that students often did not know how to act on the feedback they received, which was confirmed by Ferris (1995), who found, in an investigation of feedback given to 155 ESL university students, that 50% of the students had problems understanding the teacher's comments.

Teacher/student conferences are recommended for inprocess feedback (Goldstein & Conrad 1990, Keh 1990) and such discussions could presumably alleviate the student difficulties of not knowing how to act on their written feedback, or of not understanding their written feedback. Teachers are often advised that response during the writing process is more valuable than response when it is finished (Langer & Applebee 1987), although Warshauer Freedman (1987) reports that students valued end-of-process evaluation more. The biggest hindrance to conferencing, as well as to detailed written feedback, is the time it takes. Peer evaluation is one of the methods that has been recommended in order to overcome the problem of limited teacher time (Edge 1980; Bruton 1981; Chaudron 1984; Chimombo 1986; Keh 1990; Santos 1992; Caulk 1994; Mendonca & Johnson 1994).
Despite the fact that peer evaluation is regarded as an extra resource that is a poor substitute for teacher feedback, some researchers have found no significant difference in levels of writing improvement between students who received teacher feedback and those who received peer feedback. Robb, Ross and Shortreed (1986) compared the number of error-free t-units in narrative writing samples and found that the type of feedback students had received had made no difference, at least to their levels of accuracy. In a similar study, Chaudron (1984) investigated the effects of peer evaluation on students' revision performance, compared it with the effects of teacher evaluation and found no significant difference. This finding, however, must be viewed within the limitations of the experiment, which had only a very small sample size (14 students). He argues in favour of peer evaluation on the grounds that it provides an audience at the learner's own level of development and interest, that it has an affective benefit because peers will be tend to be more supportive than teachers and that the students will learn about writing through having to read each other's drafts. Caulk (1994) found peer feedback helpful for all but 6% of the students and reported that peer feedback was specific, while teacher feedback tended to be general, so the two types were complementary. Keh

(1990) recommended peer feedback, but noted that students needed to be trained as evaluators, a task that was noted by Mangelsdorf (1992 cited in Boughey 1997) as not an easy one.

Criticisms of peer evaluation are mainly to do with doubts about the ability of students, and particularly ESL students, to carry out such operations. Davies Samway (1993) maintained that even young children could evaluate writing, although they found it hard, but many studies found either that students hardly used peer comments in their revisions (Connor & Asenavage 1994) or that they preferred teacher feedback (Zhang 1995). Sengupta (1998) commented that the students in her study did not perceive peers as 'real readers', since their primary aim was to achieve error-free writing and wanted evaluation that would help them to achieve that. Another criticism of peer evaluation is that the practice may cause conflict and that students might be more concerned with harmony than with evaluation (Nelson & Murphy 1993; Carson & Nelson 1994).

2.5.1.3 Rewriting

Once a preliminary evaluation of the writing has occurred, whether it is made by the writer, a peer or the teacher, the writer faces the problem of whether to rewrite and how extensively to rewrite. Charles (1990) advocates getting students to annotate and mark their own essays, while Davidson and Tomic (1994) suggest that using computers can help ease the problems of revision once the evaluations have been made.

The process theory of writing that has emphasised the importance of multiple drafts, but has not translated easily into pedagogical practice, since teachers can sometimes be reluctant to intervene for fear of disempowering their students (Reid 1994). The teacher's role as evaluator is perceived by the students as necessary (Zhang 1995) and yet even when students receive the feedback they feel they need, i.e. feedback particularly on form, it seems that they are often unable to make changes that improve their writing. Sometimes repeated drafts can even make the writing worse (Beach 1979 in Leki 1990). Many researchers have commented on the fact that weak writers make only surface changes, and that these are often not an improvement. Allwright, Woodley and Allwright (1988) evaluated reformulation as a practical strategy for the teaching of academic writing and found it to

have a generally positive effect, although they stressed that practice in reformulation was more valuable as an exercise in itself than it was for improving the writing, a somewhat strange conclusion. Cohen and Cavalcanti (1990) compared NS and NNS students and found that the EFL students they investigated did not usually rewrite, whereas NS students found it easier to criticise and discard their work. This suggests that the value of rewriting may be linked to the level of writing proficiency.

2.5.2 Types of writing practice

There is consensus that practice is needed in order to improve writing skills. Robb, Ross and Shortreed (1986) investigated the development of writing competence in narrative writing by looking at five essays written at equal intervals during a year, and compared different kinds of feedback on writing. They concluded that practice was the crucial factor in promoting improvement in writing skills, not the presence or absence of feedback nor the practice of one kind of feedback as opposed to another. Leki, Ilona and Carson (1994) offered support for this view with their comment that practice improves writing far more than isolated teaching does.

There is, however, no consensus as to the type of writing practice that best enables the development of writing competence. Elbow (1991) recommends that students are given narrative writing tasks, even in academic settings, because they need to write by choice so that can get practice and learn to write well. The implication is that students would never voluntarily choose to write academic types of writing. Narrative writing practice, particularly for learning journals in teacher training for example, is widely used and its value acknowledged (Cortazzi 1994). Ross, Robb and Shortreed (1988) advocate journal writing as a means of increasing fluency, but warn that limitations become apparent when students are asked to produce more difficult types of writing. Carroll (1994) stresses that journal writing is not simply record keeping but a process of evaluation, although Stevenson and Jenkins (1994) found, in their investigation of the use of journal writing in the training of international teaching assistants, that there was little evidence of the kind of analytical, reflective writing that had been expected. They did, however, find it beneficial for stress management. Raimes (1998), too, notes the affective benefits of journal writing, although Holmes and Moulton (1995) report the case of Dang, a student, who did not

like informal writing and was not receptive to the practice. In general, however, journal writing is found to be helpful in order to encourage reflection and to improve fluency (Casanave 1994).

While teachers have traditionally regarded narrative as the best kind of writing for students to discover themselves and their thought processes, they have also regarded it as the kind of writing that students do only until they can tackle academic writing (Blanton 1994). Many have considered it inappropriate for college purposes. Blanton argued, however, that narrative writing did have a place in a college curriculum because good academic writing required authority, without which there was a sense of powerlessness, a view shared by Elbow (1991). To achieve authority in writing, Blanton argues that students should be given the opportunity to bring their own view and experience to bear on texts and that the best way to do this is give them the opportunity to produce narrative text. Belcher (1997 cited in Raimes 1998) endorsed such a view by approving a student's use of narrative to present data for a dissertation on teachers' reaction to a content-based curriculum. It seems that some researchers are crossing traditional academic boundaries in the types of writing they use and recommend.

Creative writing, however, seems to be rarely used for second language practice. The awareness of the importance of fostering the imagination of first language writers is evident as, for example, in Elsner's (1991) point that if American education is to develop minds, then imagination must be fostered, but there is an unstated assumption that second language writers will benefit most from the types of writing that are seen to be directly useful, such as letters, descriptions, reports or persuasive essays. McKay (1982), for example, saw only a very limited place for literature in the ESL classroom and argued that English literature was not suitable because it reflected a cultural bias, which she presumably considered to be a bad influence on ESL students. She argued, too, that it was not suitable because of its structural complexity and because it was not helpful for academic or occupational goals. She did concede, however, that literature might be useful if it were very carefully selected. Parisi (1979), on the other hand, advocated creative writing as a means of understanding literature and learning to appreciate it and argued that it enabled students to access more fully the culture of the target language. More recently, Raimes (1998) cites several studies which also recommend using literature in the teaching of ESL students. Morgan (1994), for example, recommended teaching poetry

writing, even for beginners, because it emphasised the shift away from product, where 'good' poems would be the aim. The focus should be on the process with the aim of getting students to roll round in the language, to make it their own, to express their emotions and to enjoy it.

The writing task affects the motivation that drives the process. Donaldson (1978 in Cummins 1983) makes the point that we can easily both underestimate and fail to elicit students' powers of language production because we fail to offer writing topics that interest them. Children operating in their L1 in school situations and ESL students in various institutions are often required by the educational context to produce the kinds of speech and text that they are not interested in. This makes an enormous difference to their levels of motivation. There has been concern, particularly in L1 writing research, over the impersonal nature of academic writing (Mlynarczyk 1991). Elbow (1991), as mentioned above, recommends that students should learn to write non-academic as well as academic discourse as a means of getting personal thinking clear. Martin (1985), on the other hand, stresses the need for students to learn how to write factually.

Attempts to teach students how to write have proceeded in a number of ways. One method has been to elicit from students a model of the type of text that they would be required to write, or to provide them with one directly. In the 1980's there was wide recommendation for students to analyse texts and then write from models (Whaley 1981; Gordon & Braun 1983; Tomlinson 1983; Stahl-Gemake & Guastello 1984). Watson (1982), however, emphasised its disadvantages. She felt that any advantages were outweighed by the negative effects of using such a method. The weightiest negative effect she described was the mechanical, artificial writing that tended to result from such methods. The other disadvantage she identified was 'false reassurance'. However, one of the difficulties of the composing process is recognised to result from the lack of an internalised discourse model caused by an inadequate amount of reading. It is students who suffer from such a lack who often appreciate the reassurance that models offer and who are at too basic a stage to worry about whether or not the reassurance is false.

Analysis of texts to discover models has since reappeared in a new guise, inspired by the insights of genre analysis, which is widely recommended as an aid to student planning and writing and as a means of showing how and why discourse types differ (Yarapawa 1991, 1994, Dudley-Evans 1994)). Swales (1990) has written influentially on how the practice of genre analysis can be a valuable approach for teaching academic and research writing. Genre analysis informs investigations into how knowledge is created in writing (Halliday 1994), as well as raising awareness of the features associated with particular discourse types. The genre-based pedagogy practised in parts of Australia is currently taught by using a teacher-led modelling phase of the text-type, followed by student practice using the model (Hyon 1996). Genre-base pedagogy, however, as mentioned above, has been criticised for its tendency towards rigid imposition of a preset form (Raimes 1998), which may impose inappropriate styles on some writers (Johns 1995). In defence of genre as a useful orientation for teaching, Purves (1991 cited in Johns 1995) makes the point that many ESL students both need and value help in absorbing and understanding text structure, a need that process writing does not meet. One of the most pressing unresolved problems, however, is the fact that the hierarchy of genres or writing types on which current syllabi are based, are only partially supported by empirical evidence. The hierarchy is intuitively appealing, but stands mostly on informed guesswork and faith.

2.5.3 Practical constraints

Whatever method or kind of writing practice is used and whatever kind of feedback is given, there are practical constraints on the teaching of writing that strongly influence decisions. The practical constraint of time available for marking is the main dictator of the number and type of writing tasks given to students (Brown 1991). Spencer, Lancaster, Rey, Benvie and McFayden (1983) reported that in most Scottish secondary schools more than half the writing tasks were copied writing and most continuous writing tasks required less than a page. Reading and marking student essays is time-consuming and teacher time is a limited resource worldwide.

Another constraint on teaching can be the student perceptions of the task. Increasing notice is being taken of student perceptions, not only because there is a genuine belief that students are worth listening to, but also because TESL has become big business and money will not be made if clients' views are

not taken into account. In the USA a process writing course for graduate NNSs did not work very well because the students did not like peer revision and collaboration, lack of a transmission model and lack of grammar teaching. This prompted the researchers to ask, firstly, whether students should have to receive content and method that they did not like, and secondly, whether ESL teachers were appropriate experts to make decisions on the teaching of writing given the different specialised subject areas of the students (Silva, Reichelt & Lax-Farr 1994).

Clarke (1994) has pointed out the frequent dysfunction between theory and practice. He gave four reasons for this: i) theorists don't practise, ii) theory is often imported from other disciplines and it doesn't fit, iii) theory is often general and does not fit specific instances, and perhaps most damningly he points out the fact that iv) theorists often underestimate the institutional, political and interpersonal constraints on putting the theory into practice. When theory does not fit the practice, then the practice goes its own way. Unfortunately this does not always feed back into the theory as a process of correction.

The review undertaken indicates that the following problems exist:

- current theory of the process of writing is difficult to teach;
- incomplete theories of how writing skills develop make syllabus design difficult;
- confusion as to what makes a good piece of writing make both teaching and evaluation difficult;
- the requirement for students to practise writing frequently is hard to achieve because essay correction is time-consuming.

CHAPTER 3 - RESEARCH DESIGN

3.1 Aims

The overall purpose of the research is to describe the development of writing competence in grade nine Papua New Guinea high school students.

The three main aims of the research are:

1. to investigate the relationship between personal history narrative, imagined story narrative and persuasive writing as produced by grade nine high school students who speak English as a second language;
2. to chart the development of writing competence over three quarters of a school year in personal history narrative, imagined story narrative and persuasive writing;
3. to investigate the effect of practice in imagined story narrative, as opposed to the effect of practice in personal history narrative, on the development of writing competence and the transition to persuasive writing.

3.2 Reasons for choice of writing types

The reasons for choosing narrative, rather than descriptive writing for example, as the type of discourse most appropriate for the investigation of the transition to persuasive writing, are as follows. Firstly, narrative was chosen because it is the easiest type of writing to produce (see Chapter 2.1.3), secondly, my subjects would already have had experience in this kind of writing, and, thirdly, I hypothesised that narrative, especially invented narrative, contained within it aspects of discourse that were needed for persuasive writing and so would provide a preparation for the transition to more academic forms of discourse. Shrubsall supports my view with reference to narrative discourse in general:

> 'Telling a story is not a discrete activity that can be left behind as we get on with other language activities, for example, arguing, reasoning, explaining. Rather, story-telling is essentially related to these other ways of using language often thought of as exhibiting 'higher order' thinking skills. When we tell stories we are often doing many things at once: for example, pursuing an argument, locating ourselves within a social space, performing an academic task.' (1997:402)

and cites Fox (1989) in support of the view that elements of invented narrative are especially helpful as a preparation for academic discourse:

> 'It may be that for a lot of the time in ordinary life children find it difficult to see viewpoints other than their own, and are not yet highly developed reasoners, explicators, or rational debaters. But in the context of a familiar set of discourses, those of fantasy storytelling, which on the face of it might be the last areas we would investigate to discover the more rational aspects of their thinking, they sometimes show embedded in the pleasure and excitement of their 'yarning on', cool and serious minds at work, weighing up the pros and cons of imaginary problems, problems posed by themselves as narrators, and solved in satisfying and elegant ways.' (1989 cited in Shrubsall 1997: 403)

Since the primary aim of the research was to discover how students might be helped to become more proficient in persuasive writing, there will be a brief discussion of the characteristics and relationships that I believed to exist between the three kinds of writing chosen for investigation. This will make clear why these categories were chosen for the research. The discussion will deal with my view of: i) the differences between narrative types of writing, ii) the hierarchy of difficulty implicit in the differing cognitive requirements of the chosen writing types, and iii) imagined story narrative as a bridge to persuasive writing.

3.2.1 Differences between narrative types

Narrative writing is not usually separated out into different types because it shares a common structure. There are, however, three clearly discernible types of narrative from the point of view of the cognitive processes involved in production, as described in Chapter 1 (see Chapter 1 pp 9-11). The three types are: Other People's Narrative (OPN), Personal History Narrative (PHN) and Imagined Story Narrative (ISN). These three types have different functions and these functions require different cognitive processes.

The term 'OPN' is used to describe all the stories that have been generated in the first place by other people. These could be myths or legends, they could be stories of films seen on television, they could be tales of other people's personal experience. They could be fiction or fact. OPN seems to be the easiest kind of narrative to produce because such narratives have already been given a structure and

need only to be retold. The main function of telling OPN is to entertain and perhaps to teach or to learn. The main cognitive process seems to be memory selection.

PHN is used to describe stories that recount the teller's past personal experience. They are factual. This type of narrative seems to be the second easiest to produce. The writer is required to select experience from memory and to express it as a chronological chain of events. The events and feelings have already been experienced but the teller needs to impose a structure so that the story makes sense. The recounting of past personal experience identifies and validates the self. The self serves as the main audience in order to come to terms with what has happened and make sense of it, while a secondary audience of others who are usually socially close also serve as validators of the experience. Considerations of audience are normally easy to fulfil because the audience is well known. The main function of PHN is to make sense of experience and secondary functions may be to entertain or to teach. The main cognitive processes seem to be memory selection, evaluating and ordering.

In contrast, ISN is a story the teller has invented. The telling of invented stories requires the writer to imagine possible experiences and to explore their consequences. The writer is required to invent a wholly new experience that has not been received, either directly or indirectly, from outside. It requires imagining some alternative reality and exploring the consequences to create a fresh experience, a new story. It offers the opportunity to explore experience otherwise denied by the limitations of external reality and widens a person's possible world from externally bound to internally unlimited. In proportion to its benefits, it requires more cognitive effort than the other two narrative types because the story has to be created, rather than simply adjusted to suit the communicative occasion. The main function of ISN is to explore possible experiences. Its function is play in the service of learning, and its secondary function is to entertain. Its main cognitive process seems to be imagining a single or integrated event chain, which requires a recurring set of mental operations: choice, hypothesis and evaluation.

OPN, PHN and ISN differ in the level of difficulty of psychological interaction between self and internalised other. An OPN is given as a gift, ready made, so little psychological effort is required to reprocess it. It can remain the story of another. A PHN requires the writer to travel into the past to make sense of experience. Some psychological interaction is required between the writer's self and internalised other, not in order to remember the events, but to order them and to provide evaluation. In contrast, the kind of experience generated by ISN, is qualitatively different because it has to be fabricated internally rather than experienced from the outside. It is not an experience that is as relentlessly part of the person as that imposed from outside by external circumstance. It is not waiting in the episodic memory bank, but has to be chosen. The choices are made, presumably, through a process of inner dialogue that includes negotiation and evaluation. The choices and the psychological interaction that the choices necessitate mean that ISN is harder to write than the other two kinds of narrative.

The psychological interaction of self with an internalised other is further affected by the differing influence of external audience, since that audience has to be imagined. We are all accustomed to retelling stories we have heard, or telling friends about things that have happened to us but we are less accustomed to telling about things that have not happened to us, and even less accustomed to telling stories or ideas to people we do not know. Consideration of external audience becomes progressively more difficult with the imposition of one or more of three operating constraints. The first concerns the increased involvement of the self and the necessary evaluation of how others will react. This means that PHN involves more audience difficulty than OPN. The second constraint is the need to perform an assessment of others' evaluation without a basis of previous personal experience. This can apply to PHN if the writer has not previously given similar accounts and is not totally sure how the audience will react. It will always apply to ISN because the story will be a trying-out on the audience of something new. The third constraint is the need to imagine how others, who are not well-known to us, will react to the discourse. This third constraint can apply to the narrative types, but is more likely to apply when writers produce formal academic types of discourse, such as persuasive writing. The constraints are powerful, not only because the audience reaction might be

difficult to predict, but also because the audience reaction might be frightening to predict. Increasing constraints of audience make writing a risky business and fear can inhibit the process.

It must be stressed that the differences that have been argued to exist between the three kinds of narrative are important for the processes involved in production. The differences refer to writing, not to the text structure of the product. This does not mean that the differences between the types of writing are negligible. On the contrary, they are substantial and must exert a powerful influence on the writing process because of the differing degrees of difficulty entailed. To distinguish between the types of narrative once the process if finished and the text has appeared, it will be necessary to be aware of the writer's background and preferably of the circumstances of production in order for the reader to judge whether the narrative was an original invention, someone else's story, or recounted from personal experience. Any differences are expected to be primarily differences of content, although the texts of immature writers may show differences in product caused by differing levels of difficulty. That the types of narrative cannot be distinguished by the outward appearance of their texts is not important because the issue under investigation is whether the different cognitive processes needed for the production of ISN help writing competence to develop. The issue concerns what processes a writer uses to produce the different types of writing, to identify differences and to discover whether practice in one type may enable the development of proficiency in another. The three narrative types share the same discourse structure but they do not share the same cognitive demands.

OPN, PHN and ISN differ, then, in their cognitive requirements. They differ in the degree of psychological interaction, or inner dialogue, required for the task and they have a differing susceptibility to audience influence. Despite the fact that they share a common discourse structure, the differences between them are substantial. The choice involved in the production of ISN and the imagination needed requires a different and seemingly more difficult process than that which is required for either of the other two narrative types.

It was decided to use only two types of narrative writing for the study: PHN and ISN. The decision was made in order to limit the research and make it more manageable. OPN seemed a reasonably easy category to exclude for the following reasons. Firstly, myths and legends are the main fictional types of OPN in Papua New Guinea and are easily recognisable. Secondly, another source of OPN would be films on television and there was only one TV channel making it easy to check which films had been on and monitor student essays accordingly. Thirdly, the main source of OPN would be the personal experience of others and it was hoped to exclude this type of OPN from the ISN category by providing essay topics that could not have been experienced personally. I realised that it would be more difficult to exclude OPN from the PHN category, but hoped that the essay topics provided would generate enough interest to motivate students to write from their own experience. In addition I thought that the most likely source of OPN would be from a student's classmates and I could control for this by keeping a careful check that students did not produce identical accounts of personal experience. PHN seemed a better choice to compare with ISN because the value of PHN as a useful starting point for writing in school was already acknowledged, and the students would have already had some experience in this kind of narrative.

3.2.2 Hierarchy of difficulty

There is a hierarchy of difficulty evident across the three types of writing chosen for the study.

Table 6: Hierarchy of difficulty of writing types according to cognitive processes

	Personal History Narrative	Imagined Story Narrative	Persuasive Writing
main function	validate self through retelling of personal experience (historian)	play with new experiences, explore possibilities & consequences, entertain others (storyteller)	change others (social manipulator)
main cog. processes	memory selection, requiring evaluation & ordering	imagination o f single or integrated event chain, requiring choice, hypothesis & evaluation	imagination & comparison of consequences of opposing ideas in a way that relates to the mindset of the intended audience
	- .....external audience.....+		
	- .....psychological interaction with self (inner dialogue).....+		
	- .....difficulty.....+		

The hierarchy of difficulty is dictated by the function of the writing. The three types of writing have different functions, as shown in Table 6 above, and these require different cognitive processes that imply differing degrees of difficulty. The heavier load exerted on STM by persuasive writing compared to narrative writing, in respect of the production of both sentence structure as well as overall text structure, was explained in Chapter 2 (see 2.2.5). For persuasive writing an additional load on STM is added by the external audience constraint, which necessitates more difficult internal dialogue than the narrative types because it has to include the imagining of the needs and probable reactions of a usually distant external other. Since the main function of the writing is to persuade, it cannot be successful without effective consideration of audience. Persuasive writing is clearly more difficult to produce than ISN, which, in turn, is more difficult to produce than PHN, as discussed above.

3.2.3 ISN as a bridge to persuasive writing

Imagined story narrative could form a bridge from PHN to persuasive writing by providing the security of a familiar discourse type while preparing the mind for persuasive writing by stretching thought in new ways. There is evidence from Shrubsall's (1997) study (discussed in Chapter 2 p 25) that the monolingual children's guided ISN oral narratives were more evaluated and more episodically structured than those of their bilingual peers. The following implication was drawn:

> 'The results have pedagogical impications for if these bilingual children are not able to use these academic (literacy/exposition -related) narrative discourse features to the same extent as their monolingual peers, they are likely to fall behind in any curriculum areas that depend upon them ... It may even be thought that story is a way into the curriculum for bilingual children who are in the process of acquiring academic English.' (1997: 414)

As mentioned in Chapter 1, students in Papua New Guinea, who use English as a second language, will rarely have been given the opportunity to write invented stories, while the practice of creative writing is taken for granted in school settings where English is spoken as a first language. I believed, for the reasons given above, that practice in writing imagined stories might facilitate the general development of writing competence and aid the transition from narrative to persuasive writing.

There are three main difficulties involved in producing persuasive writing. The first is in the effort of imagination that is required in order to explore the consequences of opposing ideas. The second difficulty consists of the effort needed to hold the two chains of consequences in mind in order to compare them. The third lies in the requirement to accurately imagine the mindset of the audience in order for the writing to succeed. These processes with their attendant difficulties have to take place in order to effect persuasive writing.

What is needed in order to prepare the mind to cope with the demands of persuasive writing might be some sort of cognitive practice that is relevant but not quite so demanding and this is just what practice in imagined story narrative might provide. Composing imagined story narratives requires choice and the exercise of imagination, and yet it is easier than persuasive writing because it requires only one main choice and one follow-through in imagination, rather than two sequences which have to be compared. Audience requirements are also less, firstly because the audience for ISN is usually closer, and secondly because one of the main functions of ISN is the exploration of new experience for the writer herself and this can be successful without consideration of an external other. In contrast, the success of persuasive writing depends on effective consideration of audience. ISN, then, is easier while sharing some mental processes in common with persuasive writing.

3.2.4 Definition of writing types

Personal History Narrative (PHN) - telling about a series of events that has been experienced personally

Imagined Story Narrative (ISN) - telling about a series of events that has been invented by the writer

Persuasive Writing (PW) - expressing ideas and giving reasons in order to persuade the reader to agree

3.3 Method

A control group and an experimental group were given a pretest in all three writing types (timed under examination conditions), a treatment of three terms, where essays were written for homework (PHN practice for the control group, ISN practice for the experimental group) and a posttest in all three writing types (timed under examination conditions).

The subjects were Papua New Guinea students of two mixed ability Grade Nine classes from Laloki High School, located just outside the capital city, Port Moresby. The original intention had been to use one class as the control group and the other class as the experimental group. This proved impossible (see Chapter 4.1), so half the students for the control group (17) and half the students for the experimental group (17) were randomly selected from each Grade Nine class. This resulted in a control group with 17 students from 9A and 17 students from 9B, and an experimental group with the same arrangement of 17 students from each class. There were 34 students in the control group (Group 1), 14 females and 20 males. There were 34 students in the experimental group (Group 2), 12 females and 22 males. A t-test was carried out to establish the similarity of performance between the two groups (level of significance set at $p<0.05$). See Table 7 below for results.

Table 7: Comparison of control group and experimental group on pretest essays

Writing Type	n	Gr1mean /15	n	Gr2 mean /15	t	p
PHN	34	9.18	34	9.00	0.32	0.75
ISN	34	9.18	34	8.74	0.88	0.38
PW	34	7.24	34	7.18	0.11	0.91

The pretest scores in all writing types showed no significant difference between the groups, so it can be assumed that the groups had similar writing proficiencies at the start of the experiment.

Pretest

There was a one hour timed pretest for each writing type: PHN, ISN and PW. Each test had three titles to control for topic effect in order to ensure that the differences in performances resulted from

the kind of writing under investigation and not from the particular topic. (See Appendix B for a list of pretest essay prompts and 3.5.1.1 below for a discussion of prompt design.)

Treatment

I taught both groups.

The Control Group (Group 1) received one lesson per week (40 minutes) and one homework per week (40 minutes) to practise PHN over a period of eight months from March to October in 1990. During this period students were expected to produce 20 essays, which they wrote, untimed, both in class and for homework. (Please see Appendix F for a list of titles.)

The Experimental Group (Group 2) received one lesson per week (40 minutes) and one homework per week (40 minutes) to practise ISN over a period of 8 months from March to October 1990. During this period students were expected to produce 20 essays, which they wrote, untimed, both in class and for homework. (Please see Appendix F for a list of titles.)

Posttest

There was a one-hour timed pretest for each writing type: PHN, ISN and PW. Like the pretest, each writing type had three titles to control for topic effect. (See Appendix C for a list of posttest essay prompts and 3.5.1.1 below for a discussion of prompt design.)

3.3.1 Relationships between writing types

3.3.1.1 Hierarchy of difficulty

It was hypothesised that a hierarchy of difficulty existed across the three writing types. The hypothesis was based on speculation about the cognitive processes required for production of the three types of writing and it was expected that the level of difficulty associated with the cognitive process required would determine the level of writing performance.

The following hypotheses were drawn up:

1. Subjects who produce a satisfactory piece of persuasive writing will produce a satisfactory piece of imagined story narrative.
2. Subjects who produce a satisfactory piece of imagined story narrative will not necessarily produce a satisfactory piece of persuasive writing.
3. Subjects who produce a satisfactory piece of imagined story narrative will produce a satisfactory piece of personal history narrative.
4. Subjects who produce a satisfactory piece of personal history narrative will not necessarily produce a satisfactory piece of imagined story narrative.

The Gutman Scale was used to test for a hierarchy of difficulty. The scripts were scored by three independent raters using holistic impression marking as described below in section 3.5.1. For the purposes of the implicational scale, the essays were divided into 'satisfactory' (10-15/15) and 'unsatisfactory' (0-9/15) because a clear division into pass/fail essays was needed.

3.3.1.2 Objective differences

Narrative writing types usually display more numerous, shorter t-units than persuasive writing. The aim was to see if the differences which have been found to exist between the writing types were evident in immature second language writers, and to find out if there were objective differences between PHN and ISN. It was expected that there would be structural differences, as measured in number and length of t-units between narrative and persuasive writing, but not between PHN and ISN. Differences in fluency were expected to reflect the predicted hierarchy of difficulty, where students would write most fluently for personal history narratives, less fluently for imagined story narratives and least fluently for persuasive writing. Levels of accuracy were expected to differ to reflect the predicted hierarchy of difficulty. It was expected that students would write most accurately for personal history narrative, less accurately for imagined story narrative and least accurately for persuasive writing. The objective measures of structure, fluency and accuracy (described below in 3.5.2) were used to compare performance in the three writing types.

3.3.1.3 Indicators of text Quality

The aim was to discover which objective measures of structure, fluency and accuracy (described in 3.5.2) discriminated significantly between 'good' and 'poor' pieces of writing and to find out which of these were common to all three kinds of writing and which of them differed. Following work by Halliday and Hasan (1976) and others, it was expected that the objective measures that would discriminate between 'good' and 'poor' pieces of persuasive writing, yet not between 'good' and 'poor' pieces of narrative writing, would be lexical density, manifested as significantly longer t-units for persuasive writing. Satisfactory performance in narrative writing, on the other hand, was expected to display a more verbal style, so it was expected that the number of t-units would be a distinguishing feature. The fluency measure was expected to discriminate between 'good' and 'poor' pieces of writing in all types. Accuracy measures were expected to discriminate between 'good' and 'poor' pieces of writing of all types. According to research by Perkins (1980) the level of error discriminated between satisfactory and unsatisfactory writing of all kinds to the extent where he concluded that only error-free measures were significant. The pretest scripts were scored by holistic impression marking (as described in 3.5.1). For the purpose of looking at those features that discriminated between 'good' and 'poor' essays, it was intended that scores of 11-15 would qualify as 'good' scripts, while scores of 0-5 would be identified as 'poor' scripts. Unfortunately these divisions did not yield sufficient scripts to make comparison worthwhile, so scores of 10-15 were used to identify 'good' scripts, while scores of 0-6 were used for 'poor' scripts. These scripts were then analysed according to the objective measures described in 3.5.2.

3.3.2 Development of writing competence

3.3.2.1 Overall development

It was expected that the all students' writing would improve to some extent. The students' narrative writing was expected to improve more than their persuasive writing, since they had been given practice in narrative writing and not in persuasive writing. The pretest and posttest scripts were scored by holistic impression marking (as described in 3.5.1) and the means were compared.

3.3.2.2 Change in objective features

It was expected that those features of text that had been shown to be associated with quality would increase as competence developed. Objective measures, described in 3.5.2, were compared between pretests and posttests in each writing type.

3.3.3 Effect of practice in ISN on transition to persuasive writing

3.3.3.1 Overall improvement in persuasive writing

It was hypothesised that practice in ISN would have a beneficial effect on the production of persuasive writing because this writing type was believed to form a bridge between PHN and persuasive writing, as argued above in section 3.2. It was expected that the benefit would be evident through a greater improvement in performance in persuasive writing than that achieved by the control group. The following hypothesis was drawn up:

5. Practice in writing imagined story narrative is associated with an improvement in overall performance in persuasive writing; this improvement is significantly greater than any improvement in persuasive writing associated with practice in personal history narrative.

To test Hypothesis 5 the change between persuasive writing pre- and posttest scores was calculated for each student and then the groups were compared by means of a t-test.

3.3.3.2 Change in objective features of persuasive writing

It was believed that practice in ISN would benefit the development of competence in persuasive writing. The kind of cognitive practice it seemed to imply was believed to promote an automatising of the kind of mental operations needed for the development of academic writing skills. It was expected, therefore, that the experimental group's persuasive writing would show a greater increase in those objective measures shown to be associated with quality. In particular it was expected that the experimental group's improvement in persuasive writing would be shown by a significantly greater increase in accuracy due to an increased automatisation of lower level skills reducing the load on STM.

The following hypothesis was drawn up to test this expectation:

6. Practice in imagined story narrative is associated with a decrease in the number of errors in persuasive writing; this decrease is much greater than any decrease in number of errors associated with practice in personal history narrative.

To test Hypothesis 6 the difference in number of errors for each group between the pretest and the posttest was calculated and then the groups were compared by means of a t-test.

3.4 Subjects

The subjects were approximately 16 years old. All the subjects were in their ninth year of schooling in English. Almost all the students came from the Papuan coast, where the first language is Motu. Most had learned English as a second language for educational purposes and Tok Pisin as a third language for purposes of wider cross-cultural communication with people from other parts of PNG. All the students in the study spoke English and Tok Pisin and a few had a first language other than Motu. Some spoke fourth and fifth languages, because they had transferred from another area, or because one of the parents had married into the area and had brought another language into the family.

Personal history narrative, expository and persuasive writing were taught in both community and high schools, but imagined story narrative has until recently occupied a very minor role in the syllabus. It appears only rarely in the main teaching textbooks *Create & Communicate* Books 1 and 2 (Heaton & The Papua New Guinea Dept of Education 1985, 1986, 1987, 1988) that the students had used so far in their secondary education. It is unlikely that the subjects had received much practice in imagined story narrative prior to the experiment. Their main source of reading material was the *Post Courier* as described in Chapter 1 p3), which they read for pleasure whenever they could get hold of a copy. The main source of reading material should have been the school library, which was excellent by PNG standards. The number of attractive up-to-date books in Laloki High School's library was a tribute to the vision of successive headteachers as well as to the hard work of the staff

and students since they had worked very hard to earn money to buy books through projects such as bee-keeping and honey production. (*Laloki Honey* was excellent and much sought after.) Unfortunately, students had so much physical work to do in addition to their study for other subjects, that there was little time for reading. They would, however, have read some books, especially graded readers. For 1990, the year of the study, the writing sections of the English lessons programmed for the Grade Nine syllabus were replaced by the writing project designed for the experiment. Other English lessons were taught mainly from the Grade Nine textbook *Create and Communicate* Book 3, which included intensive reading texts that were often experienced as extremely difficult.

Laloki High School is situated just outside the capital city of Port Moresby and in many respects can be considered part of the capital. Although many students in the sample were not originally from Port Moresby, they were all exposed to the urban environment at the time of the study. The urban environment meant an increased presence of English, both aurally and on notices etc., compared with the amount of English found in rural areas. The point is being made that although Papua New Guinea is still almost totally rural, the sample was taken from an urban environment, and this factor can be assumed to have affected the results.

3.5 Evaluation of Essays

The essays were evaluated in two ways: i) by holistic impression marking, and ii) by objective measures.

3.5.1 Holistic impression marking

In order to assess the overall quality of writing, the scripts were scored by holistic impression marking. This is the kind of text evaluation we do naturally as readers. We read and absorb the text as a whole. We react to its content as well as its style and come to a decision about whether or not we thought it was interesting and well written. It is the kind of evaluation that is considered to be the best (Huot 1990) because there is still no accepted theory of writing that has identified all the features which contribute to a text and their interaction with each other and relative importance. For this

reason it is the kind of evaluation that has the highest face validity (Stansfield 1986) and it was the lack of face validity which influenced TOEFL to abandon its objective writing tests in favour of the Test of Written English, which tests writing directly by subjective evaluation regulated by detailed criteria. IELTS also uses similar methods of evaluation.

Direct writing tests that use subjective impression marking, however, have problems of standardisation and reliability. Three main areas have been identified that need to be controlled in order to achieve reliability: i) the test prompt, ii) the rating scale, and iii) the raters.

3.5.1.1 Test prompts

In order to control for topic effect on performance, three titles were provided for the pretest and three for the posttest. Titles were distributed randomly so that approximately equal numbers of essays were written for each title. To achieve reliability it is necessary to control for discourse type, topic familiarity and audience specifications (Hamp-Lyons 1991e). Care was taken in these areas while designing the essay prompts for the pretests and the posttests (see Tables 7 and 8 above) and all the prompts were checked by three PNG high school teachers who thought they were suitable. The primary concern was to elicit the desired writing type, since the main purpose of the experiment was to compare performance in the different types.

Care was taken with the PHN prompts to ask for experience that all students could write about. The topic content for the prompts was familiar for PNG students. Previous experience had shown that students enjoyed writing about their celebrations so these were chosen as pretest topics. The posttest topics were based on my knowledge of experiences that were important to PNG high school students. The first schoolfriend changes your life when you are lonely and homesick, as most PNG students are when they leave home to go to school. For the same reason, the arrival of the mail (which might only come once every month or so in some areas) was eagerly awaited and sooner or later everyone would receive a present from home. My subjects attended school in an urban area, so there were more day students than is usual in PNG secondary schools, but all the students would

have received a present at some point, something to cheer them and keep them going through the difficulties of school life. The worst punishment topic was also an experience that was close to the heart of every student. Punishments were given for failure to cut enough grass or turn up to cook the rice for breakfast. Both day-students and boarding students were required to do 'workparade' duties and everyone failed sometimes and was punished. Audience consideration was expected to be determined by the writing purpose although students were aware that teachers outside the school would evaluate their test essays.

For ISN prompts the prime concern was to stimulate the imagination of the students and motivate them to invent stories. To choose a topic that ensured familiarity would be at odds with the writing purpose. I did, however, try to include elements of familiar situations in order to provide starting points for invention. For example, my subjects were familiar with the lives of birds, fish and pigs, so I hoped their knowledge would provide inspiration to write the 'Day in the life of...' essays given for the pretests. The posttest prompts were wish-fulfillment titles, based on listening to many high school students during the four years I spent as a high school teacher in PNG. Since students frequently felt lonely and homesick, I believed they would enjoy inventing a secret friend to talk to. Nearly all students prayed for something exciting to arrive in the mail, as mentioned above, so I thought that they would enjoy writing about an unusual present. Another frequently recounted high school experience was of trusting someone, who subsequently 'tricked' you and the consequent pain and anger this generated, so the third posttest prompt offered the chance for students to imagine an invented punishment for a betrayer. I put the familiar experience into a new setting ('You are Queen (or King) of a large country.....') so that students would have to use their imagination and so they could be freed from the reality of having to behave as they thought they should. As with the PHN prompts, audience was not specifically stated as I believed the writing type would determine the level of audience consideration.

For the persuasive writing prompts the primary concern was again that the prompts would generate the type of writing needed for the study. The second concern was to ensure not only that topics would

be familiar, but that they would be interesting. Both pretest and posttest topics required students to write about issues that were being debated at the time. The desire to ensure that the topics chosen were of current interest meant that the pretest topics and posttest topics for persuasive writing were not designed at the same time, unlike the pre and posttest prompts for PHN and ISN. At the beginning of 1990 PNG was becoming increasingly violent. There were increasing numbers of people moving to urban areas in hope of a better life, but who had no jobs and no land for growing food. In order to survive some of these people joined the rapidly increasing numbers of marauding gangs, who, in rage and frustration, not only robbed, but raped and killed at random. It was suggested that film violence was promoting the growth of such behaviour and censorship was a hot issue. This provided the topic for the first pretest title. The contribution of alcohol to violence was also much discussed and some provinces were experimenting with laws banning the sale of alcohol. This provided the topic for the third pretest topic. The other pretest topic was the issue of whether or not people should be fined for littering. This topic was a national issue, but was discussed particularly in the school context since schools tried to impose western-type standards of tidiness that were frequently perceived as alien. Audience for these essays was not stated specifically, but it was expected that students would write as people wrote to the *Post Courier* since this was their main model of persuasive writing and letters to the *Post Courier* were frequently used in school for both teaching and testing purposes.

The persuasive writing posttest prompts were devised towards the end of 1990 shortly before they were needed. By this time a new and highly contentious solution to the problem of urban drift had been suggested: that people should no longer have freedom of movement and that only people with jobs or means of support should be allowed to live in urban areas. This provided the first posttest topic. The alcohol problem had acquired a new focus and was now being debated in relation to the introduction of new road safety laws, which were perceived by many as an intrusion into their freedom for no good reason. This provided the third posttest topic. The other issue was the right to choose a marriage partner. Traditionally marriage was arranged and required new husbands to make large bride-price payments to the brides' families. This issue had been debated off and on for a

number of years, but the growing awareness that sexual frustration caused by inability to raise the money to get married was contributing to the increasing incidence of sexual violence, sparked the debate with new intensity. For the posttest topics students were specifically instructed to address their essays to *Post Courier* readers. (Please see Chapter 6 for a discussion of the effect of test prompts.)

3.5.1.2 Holistic impression rating scale

In order to standardise scoring of scripts, a scale was drawn up which ran from 0 (very poor) to 5 (excellent). The kind of writing that was intended to characterise each scale step was carefully described in the rating scale under three headings: 'organisation & clarity', 'interest', and 'accuracy'. The category of 'organisation & clarity' was intended to focus rater attention on rhetorical features of text organisation and for the persuasive writing on logical development of ideas. The term 'clarity' was preferred to 'logical development' since the scale was intended for use with all three writing types and narrative does not always demonstrate development that could be described as logical. 'Interest' was preferred to the alternative description of 'content' because I was afraid that 'content' might be viewed as an instruction to evaluate for 'accuracy of content'. The 'accuracy' category was straightforward since the descriptors, as well as the pre-rating discussions of the scale with raters, made clear that accuracy of language was the required focus. I discussed the scale with each rater before evaluations began, and the raters felt that the scale reflected common sense. (See Appendix B for details of the Holistic Impression Rating Scale.)

A holistic impression scale was used rather than an analytical scale, which involves the separation of the various features of a composition and requires a grade to be allocated for each feature. An analytical scale can be useful for teaching where the teacher wishes to focus on a particular aspect of the student's writing. For the purpose of this research, however, it was necessary to see which objective features appeared to indicate quality in various writing types, and so a holistic assessment was required against which such features could be matched. Holistic assessment was needed, too, in order to determine any hierarchy of difficulty between the writing types. A scoring method was needed which would allocate a grade for overall proficiency so that the pieces of writing could be

divided into 'satisfactory' and 'unsatisfactory'. Primary trait scoring was not used because a single measure was needed to assess all three types of writing. The basis of primary trait scoring is an abstraction of features peculiar to a particular discourse type in order that the particular discourse type may be considered separately from others. (For discussion of scoring methods see Chapter 2.3.)

The main problem with holistic scoring is that of reliability and so every care was taken to put recommendations for reducing subjectivity into practice (Perkins 1983). This involved providing three titles for each test to control for the possibility of performance varying according to topic, as described above. It involved development of the behaviour-specific rating scale discussed in this section. To achieve reliability it was also recommended that at least three independent experienced raters be used.

3.5.1.3 Raters

Each script was marked by three independent raters. None knew the mark assigned by the other two and none knew the students in the sample. Each rater gave a mark out of 5 in accordance with the scoring scale. The scores for each essay were then added together to give a mark out of 15. Since doubts have been expressed as to the validity of ratings when raters are exposed to pressure to agree with each other (for example, Wiegle 1994), no discussion took place between the raters and no attempt was made to standardise ratings. This was deliberate in order to ensure that each rater's perception was valued and that no pressure was exerted for a rating to be changed just because it was different from the rating of another. (See Chapter 2.3.2.4.)

The raters were experienced high school teachers who were familiar with Grade Nine level student writing. In addition, all three had experience in marking for the national Grade Ten examinations, so they had extensive experience of high school writing and the standard that was expected. They had been considered reliable markers by the Measurement Services Unit of the National Department of Education, who had employed them for the marking of national examinations. One of the raters was

a Papua New Guinea non-native speaker male teacher, and the other two raters were expatriate native-speaker female teachers.

3.5.2 Objective measures

Syntactic maturity, fluency and accuracy were objective measures chosen for the study, partly because they had been identified by previous researchers (see Chapter 2.4) as indicators of text quality and writing development, and partly because they were issues of concern to teachers. These objective features were used to investigate which features were associated with evaluations of quality and to find out which features changed as writing developed. The objective measures used in the study are listed in Table 8 below.

Table 8: Objective Measures

number of words per essay	fluency
number of t-units* per 100 words number of words per t-unit* number of error-free t-units* per 100 words number of words per error-free t-unit*	grammatical structure
number of errors per 100 words	accuracy *Categories of error:*** Vocabulary Grammar Cohesion & coherence Spelling Other

*Definition of t-unit: A t-unit or minimal terminable unit is defined as 'the shortest unit ..... which a sentence can be reduced to, and consisting of one independent clause together with whatever dependent clauses are attached to it.' (definition taken from Longman Dictionary of Applied Linguistics, 1985:299-300)

** Please see Appendix E for descriptions of error types and examples.

Fluency has been widely identified as an indicator of writing growth or quality, for example by Ferris (1994). All the pre and posttest essay prompts contained the instructions: *Write as much as you can in the time available. Time - one hour.* The instruction to write as much as possible in the time available was intended to signal to the students that fluency was important. The essays were written under test conditions.

Syntactic maturity indicated by the number and length of t-units have been widely identified as signifying writing growth, as, for example, in the attempt towards an ESL development index by Larsen-Freeman and Strom (1977) and Larsen-Freeman (1978). In connection with these measures Perkins (1980) correlated objective measures with holistic ratings and concluded that error-free measures were the only ones that discriminated among holistic evaluations. I decided to include error-free t-unit measures because of this finding. Measures of t-units that were not error-free were taken, firstly in order to test Perkins's findings, and secondly because there is an issue of a PNG variety of English. How PNG English should be described and whether it should be formally accepted is still a matter for debate, but its existence is a fact (Barron 1986; Hyland 1990). In PNG English some forms of language are acceptable that would be considered errors if counted under a typology which uses standard British English as a yardstick.

Measures of accuracy were chosen for two reasons. The first reason was because of the general consensus of research findings that accuracy is a significant indicator of ESL writing proficiency, for example Perkins (1983). (See Chapter 2.4.2 for discussion.) The second is the concern that teachers in PNG show for accuracy. There is a feeling that accuracy is important, but to my knowledge there are no PNG studies available to indicate to what extent or at what level this might be so, nor to indicate which types of error might be more important than others. Two error analyses have been carried out to determine the most frequent kinds of error and to speculate on their causes. The first of these was carried out by Smithies and Holzknecht (1981) when they investigated common errors of students at the PNG University of Technology. The second was carried out by myself (Phillip 1986) when I investigated the communication skills of public servants. Smithies and Holzknecht (1981) found that the most frequent errors were with articles, prepositions, verbs, nouns and spelling in that order. They found relatively few errors of adverbs, adjectives, pronouns, word order or style. I found similarly with the exception of the style category, which caused problems in the writing of public servants because they liked to use 'formalese' in order to impress with the result that sentences sounded important but became incomprehensible.

I used a combination of Smithies and Holzknecht's (1981) error categories and my own (Phillip 1986) as a starting point for deciding on the error categories for this study. A preliminary analysis of data was carried out in order to test the categories, but they seemed unsatisfactory, not because they did not 'work', but because all I achieved was a long, relatively meaningless list of frequency of error in various types. I decided it would be useful to put the error types into general overall categories (see Table 8 above). My concern was not only with the reasons for error, but also with the effect of errors on the text. Bamberg (1983) had drawn attention to the difference between local errors and global errors in their effect on the text and I wanted my error categories to help make transparent this distinction. There was no easy way to describe distant cross-textual problems in an error analysis, but it seemed that some kinds of relatively local error had a greater effect than others in impeding the ease of text processing. These were cases of wrong or missing reference or cases where main verbs had been omitted. When two items do not fit together or one of them is missing to the extent that the second does not make sense, then the reader has to stop and engage in a mental process of text correction in order to continue reading. Maybe readers automatically 'correct' as they read and if they do, then a combination of items that does not make sense takes more time to adjust than a single item. I decided that errors of the former kind belonged to the category of 'cohesion and coherence'. I included problems of wrong or omitted reference in this category rather than in the 'grammar' category. I also included errors of omission where the item omitted was crucial to the sense of the text. For example, I included errors where the main verb had been omitted in the 'cohesion and coherence' category, but omissions that did not affect the sense of the text, for example, omission of articles, in the 'grammar' category. I made decisions between 'punctuation' errors and 'omission' errors in the 'cohesion and coherence' category on the basis of how the problem was most easily remedied, for example if the problem of a missing main verb could be solved by changing the punctuation, the error was classified as a 'punctuation' error, but if not, it was classified as an 'omission' error .

In summary, when students made mistakes because they did not know the meaning of words, these were counted as 'vocabulary' errors. Errors in the 'grammar' category included all those traditionally

included in such a category except for the kinds of error that seemed to have special impact on cohesion and coherence e.g. confused or omitted referent. The 'cohesion and coherence' set of errors included reference, omission, logic, punctuation and conjunctions. The category of 'spelling' errors was fairly straightforward, although doubts sometimes arose when choosing between whether to categorise an error as a vocabulary error where the meaning was not known, or whether it was simply a spelling mistake. The category of 'other' errors was a ragbag and included errors that did not seem to fit elsewhere, for example errors with style. The most substantial sub-category included under 'other' errors were errors of carelessness. These were generally errors such as 'the' instead of 'they', for example, and the category also included any errors with words that had previously been used correctly. (See Appendix E for a list of error categories with types and examples.)

Categories of error are easy to describe in theory, but difficult to apply in practice (Bartholomae 1980; Rifkin & Roberts 1995). There were many occasions during the analysis of the scripts when I was unsure as to which category an error belonged and no doubt some wrong decisions were made. Note must be made of these difficulties when interpreting the findings. Errors were identified according to standard British English as described by Quirk, Greenbaum, Leech, and Svartvik (1972). The only exceptions were common lexical items that are part of PNG English, e.g. bilum (string bag) and kaukau (sweet potato), which were not counted as errors. An error was counted each time it occurred.

PART 2: DESCRIPTION OF THE RESEARCH

CHAPTER 4 - CONDUCT OF THE EXPERIMENT

The experiment was carried out at Laloki High School, which is in the Central Province of Papua New Guinea. Before the 'writing project' (the study) started, assignments had tended to be kept to a minimum because marking was so time-consuming and the school had been, and still was, understaffed. Already overburdened teachers were taking on extra duties that sometimes involved teaching two classes at once and in addition there were boarding duties. A teacher's day, like a boarding student's day, would start at 5 a.m. with cleaning duties, cooking breakfast and cleaning up. Lessons ran from 7.30 a.m. until 2 p.m. with a break for lunch. After lessons finished, the teachers were required to supervise the students on 'workparade' which involved duties like cutting grass, cleaning the school buildings, growing food and cooking it. After supper, there would be a period of night study from 7 - 8.30 p.m. and then supervision of bedtime and lights out. At approximately 9.30 p.m. a teacher was free to have a private life, while the students' lives were so full of work that they collapsed in exhaustion and slept.

4.1 Setting up the experiment

Setting up the experiment proved harder than anticipated. Papua New Guinea is a favourite place for language researchers because the country has 869 languages. There are stringent controls in place to monitor research activity, but gaining the approval of the University of Papua New Guinea and the Department of Education proved straightforward. Using the original research design, however, ran into problems before the experiment could start. I had envisaged teaching parallel classes, where one class was given practice in PHN and the other class practice in ISN. I needed parallel classes who shared the same English teacher, but this proved impossible.

The Headmaster of the school, Mr Fauma, explained that the Department of Education had introduced a policy that prevented classes in a single grade being taught by the same teacher. This was to guard against lazy teachers, who might not cover the syllabus if they were setting work and examinations for the whole grade. This meant that the research design had to be adjusted because any differences between the groups at the end of the treatment could have been ascribed to the fact that they had had different teachers for their other English lessons. For this reason I decided to split

each of the two grade nine classes, so that I had half a group of control subjects and half a group of experimental subjects in each class. It seemed at first that such a design might be even better than the original plan, but in practice this arrangement caused considerable frustration.

I was not able to teach the control group and the experimental group at separate times because of timetabling problems. This meant that I saw 9B students at 8.00 - 8.40 a.m., followed by 9A students from 8.40 - 9.20 a.m. each Wednesday. After the first four introductory lessons where all the students were taught together, each class separated into their previously allocated control or experimental groups, labelled Group 1 (control group) and Group 2 (experimental group), so I had two separate groups in each class. The experiment was conducted with no whole class teaching, apart from a few minutes at the beginning of each session to go over language problems that the students had in common. The writing process for the treatment consisted of seven main components:

1. written essay title stimulus, provided on separate essay sheets for each group;
2. peer group writing preparation done in pairs;
3. drafting;
4. peer evaluation and feedback;
5. teacher comments written on the essay;
6. rewriting;
7. individual conferencing with the teacher once every three or four weeks.

This arrangement was advantageous in some ways, and frustrating in others. The advantages were:

- the effect of separate class treatment experience attributable to being in 9A or 9B was eliminated;
- the possibility of different teaching given by the researcher to the two groups was minimised (each group could see the amount of time and enthusiasm given to the other group, and any whole-class teaching was to do with language problems and was given to both groups simultaneously).

The disadvantages were:

- whole group discussion of writing stimulus was prevented, so I could not help with this;
- whole group discussion of problems encountered while writing the week's essay was prevented.

Before the project started I talked to the staff who taught English in the school. I explained that I hoped to give one group practice in PHN and the other group practice in ISN. The aim was to see if practice in a particular kind of narrative helped the students make the step from narrative to persuasive, or argumentative, writing with greater ease. I did not offer the information that I thought one kind would help more than the other. The English staff, under the leadership of Mr Francis Rohus, were happy to support the project. They provided me with programmes of English teaching that they hoped to carry out, but explained that they would leave the writing practice to me, because there would not be enough time for the students to produce two sets of writing. The English work that was most directly related to writing was work on connectors. The English staff dropped in on the lessons from time to time, but were not able to participate further because of their heavy workloads.

The next task was to split the classes. This was done by dividing the students on each class list, in a random manner, into two groups. Students were allocated to Group 1, the control group, or Group 2, the experimental group, before the project started, and before the researcher met the students, although the students did not actually split into separate groups until after the introductory sessions.

The pretest was administered by the school staff before the project started. Each student was required to write three essays: one in PHN, one in ISN, and one in persuasive writing. Each type of writing had three titles and these were allocated randomly by the school staff who administered the test. (See Appendix B for list of pretest titles.)

There are four terms in the PNG school year. The project started about half way through term 1 and ended about half way through term 4 (March - November 1990). The students were told which group they were in and that from term 2 onwards, each group would have different writing tasks. They

were told that the level of work was the same, but that the essay titles would be different. Students were also told that they would be expected to write a certain number of essays, but that, unlike the normal pratice for homework assignments, there would be no punishment given for failure to write the essays. The project was allocated one class lesson for each class (40 minutes) and one homework allocation (40 minutes) each week until the end of the project. In practice the students reported that they spent much more than 40 minutes each week on the homework.

4.2 Introductory lessons

There were four introductory lessons on how to write essays. These were the same for both classes. Their purpose was twofold: firstly to get to know the students, and secondly to teach a procedure for writing essays which could be used throughout the project. It was important to establish a procedure for writing the essays since the students would have to work alone for much of the time. For example, preparation for writing each essay would have to take place without teacher help or intervention, and peer feedback would often take place during night study after the essay was first written and before it was ready to hand in. The undertaking of the writing preparation stage without teacher help, apart from the title stimulus, was the main difference from what might be expected of conventional writing lessons. The possible advantage of using such a method was that it might foster student independence.

In the introductory lessons the students practised techniques that it was hoped they would use independently when the classes were split into different groups for the experiment. Brainstorming, discussing and writing down ideas, ordering information and writing outline notes were practised as a whole class, and then students practised the same methods in pairs to give them practice in the way they would have to work throughout the project. The students followed up these initial steps by writing an essay. They were instructed not to worry if their plan did not match their essay. This was in accordance with the view that inflexible plans could hinder, rather than help, essay writing (Rose 1984). Finally, students were taught methods of evaluation. They were asked to read their own essays, firstly to check whether the essays made sense and were coherent, and secondly to proofread

for errors. After carrying out their own evaluations they were asked to give the essay to a neighbour to get comments that might be helpful. The teacher feedback came last, after the process of self evaluation followed by peer evaluation. After written comments from the teacher, the student was expected to rewrite the essay to improve it.

The introductory lessons were useful in several respects. The essay writing method described above was successfully established and followed through in the writing project, although the evaluation sessions that took place sometimes suffered from lack of time. The researcher and the students got to know each other a little and mutual respect was established. We felt that the writing project was worthwhile. In addition, three problems emerged. The problems were:

1. 9B students, who were taught during the first lesson of the day, were sometimes late because assembly had run over time, or because they were involved in start-of-day duties that had to be completed before they could come to class;
2. 9A students seemed to take about five minutes to move from their normal class places to their group places after the end of their first lesson;
3. the students did not have any rough paper for drafting and were afraid of putting pen to paper in case their writing might not look neat.

The first two problems were never solved and shortage of time continued to cause frustration throughout the project. I attempted a solution to the third problem by assuring the students that I was happy to see rough work and crossings-out in their exercise books, but the students merely smiled and continued to produce only neat work. My second attempt to solve the problem was through an approach to the headmaster to request an additional exercise book for each student, so that rough work and neat work could be kept separate. At the same time I went begging for scrap paper round the offices of UPNG. This was successful but time-consuming. A total of 68 students took part in the project and this meant that a great deal of scrap paper was needed. The Headmaster eventually managed to provide an additional exercise book for each student and these ended up being used for the essay rewrites.

4.3 Pattern of Progress

The number of essays written during the early stages of the project was far fewer than at later stages. There were 20 titles given for each group. Students were told that they were expected to write all 20 essays. There was no choice of title. This worked out at roughly one essay per week not counting holiday times. In addition there were 3 optional titles for each group in case some students wished to write more. I made it clear that although I hoped that students would produce the essays in a regular fashion on a weekly basis, that they could 'catch up' at any stage and produce missed essays at any time up until the end of the writing project. (See Appendix F for a list of treatment titles and Appendix G for a sample treatment essay.)

No punishment was given for failure to produce the weekly essay, as mentioned above, but students who had not produced an essay were asked for their reasons. These included the usual tales of sickness, or strange and disastrous happenings to the essays which had certainly been written but which could not be handed in because they had been swallowed by snakes, chewed by dogs, stolen by other students and on one occasion whisked away by spirits to another world. This practice in itself must have provided some exercise in imagination. My request for explanations obviously put pressure on students to complete the essays, although a few students managed to resist this pressure remarkably well.

I exerted an additional pressure on the students for completion of essays by informing them that a report on their progress, including number of essays written and marks awarded, would be made available to their parents at the end of Term 2 and again at the end of the project. The announcement of the forthcoming reports seemed to be more effective in motivating the students to produce than requests for information on their failure to do so. Table 9 below shows the pattern of essay production.

Table 9: Essay production

	Control Group (PHN) av.no of essays		Experimental Group (ISN) av.no of essays	
Term 2 (total of 9 essays)	6.44	71.6%	6.56	72.9%
Total (total of 20 essays)	16.19	81%	16.5	82.5%

By the end of Term 2 almost three quarters of the students were producing essays each week, although this proportion had built up gradually from a very poor response at the beginning. From the information given in the questionnaires (reported in Chapter 5) it became clear that students had found the essay writing very difficult but it seemed that practice gradually helped fluency. By the end of the writing project the proportion of essays produced had reached over 80%.

Each essay was given a mark out of 25 and there was little difference between the groups in average performance. There was a dramatic improvement in performance between the first half of the project and the second half. The average marks of both groups increased by more than half. See Table 10 below.

Table 10: Performance on treatment essays

	Control Group /25	Experimental Group /25
Essays 1 - 9	12.15	12.04
Essays 10 - 20	17.02	17.00
Overall average	14.59	14.52

During the first half of the writing project, everything seemed to be running smoothly and students from each group appeared to be producing the kind of essays that were required. Towards the end of the project, it became apparent that some unexpected developments were in progress. One or two of the control group students, who were supposed to be writing essays about personal experience, were quite obviously writing imagined story narratives. One student in the control group twice produced two essays: a PHN essay and an ISN essay in addition. I praised the student for producing additional

essays and marked both essays on each occasion while noting that unforeseen developments were taking place.

It seemed at that point that the only crossovers in writing types had been from the control group to the experimental group, i.e. from PHN to ISN, and not in the opposite direction. It also seemed as though the crossovers were isolated and few in number. They were, however, unexpected so I prepared two questionnaires to check what kind of writing types the students had actually written. Students were asked to categorise the type of experience they had used in each essay. It seems from the data contained in the questionnaires that although there had been much more crossover from PHN to ISN, there had in fact been crossovers in both directions. The data from these questionnaires are reported and discussed in the next chapter. (See Appendices F and H for details of questionnaires.)

4.4 Feedback

4.4.1 Peer feedback

Peer feedback in classtime took the form of oral comments by the student's neighbour. It seems likely that peer feedback at other times took the same form. Despite instructions, there was not much evaluation contained in this form of feedback apart from expressions of liking the neighbour's essay. What the feedback did seem to provoke was discussion of the content of the essay. For example, students often compared experiences and added to the information in the essays. In other words they started 'telling stories'.

I observed what was happening, listened to, and sometimes joined in discussions while walking round the class. There are two possible reasons for the lack of evaluation by peers. One reason could be that the students are not capable of useful evaluation, although many research findings contradict such a notion (e.g. Davies Samway 1993). A second reason is cultural and this seems the more likely explanation. It is an alien concept for Melanesian people to criticise each other's work. Criticism from a peer is usually regarded as negative and rude, so I did not try to force the students to go

further in their evaluation, accepting that what actually happened might be useful in itself. It is difficult to assess the contribution of peer feedback to the process. From time to time I encouraged the students to talk less and get down to writing, since it was easier to talk than to write and time was short.

4.4.2 Teacher feedback

While the students planned, wrote and rewrote essays, I talked to students one by one at the front of the class, where they came to sit down and discuss their essays. I tried hard to be as fair as possible in allocating each student equal amounts of time, but in practice I saw those students who had failed to hand in essays and those students who had obvious problems more frequently than those students who were doing well. Care was taken to devote equal amounts of attention to each group.

Discussion of the essays was mostly initiated by myself. The students rarely asked questions or brought up specific points to discuss, although they had been encouraged to do so. I think this would have been different if I had had more time to get to know them, but just over an hour a week for 68 students did not allow enough time. The conferences mainly took the form of discussion on content, which I hoped would be a motivating factor since every writer likes a positive response to their writing and respectful consideration of what he or she has to say. It is true that Fathman and Whalley (1990) investigated the relative merits of feedback on content versus feedback on form and found that feedback on grammar improved accuracy more than feedback on content improved content, but the precise effect of feedback is difficult to quantify. It seems that any motivating factor will be useful. The discussions on form that did take place were mainly to do with strange or confusing links between sentences or sometimes between paragraphs. Grammar was rarely discussed, although grammatical errors were often pointed out in my written comments. The emphasis on content and overall essay organisation was deliberate since these seemed to be the more important aspects of essay writing.

When giving written feedback I tried to pick out some points of grammar, spelling or punctuation as well as making a comment on content at the end of the essay. Sometimes comments were made on the organisation of the essay, and such comments tended also to appear at the end of the essay. Some essays inevitably received more feedback than others. Comments on content varied from very specific comments such as '*How wonderful that you got your shorts back, Willie!*' (in response to the essay 'A Friend in Need') to vague, but hopefully encouraging comments such as 'interesting ideas' or 'a nice essay'. Comments on content were specific as often as possible in order to show the students that what they had written had been considered with interest. The point is made repeatedly that it is important to take content into account and that very often this is not done in the case of ESL students, where accuracy seems to be the overriding concern (Raimes 1979; Cohen & Calvacanti 1990; Intaraprawat & Steffensen 1995). I was once told of an extreme example of such a one-sided focus. Apparently a young girl in her grade seven year had written in an essay '*My aunty die last week... '* and the teacher had written underneath '*My aunty died last week - tense!*'

4.5 Rewriting

The effect of the feedback is difficult to assess. The surface corrections were sometimes, but not always, implemented in the essay rewrites, although the same problems would recur either in a different place or in the next essay. The observations made in this research seem to confirm the findings of Brewer (1988), and Cohen and Cavalcanti (1990), that feedback on accuracy is not helpful, although other research findings contradict this. It may be that conscious awareness of grammar problems may be helpful in an indirect way after a period of gestation. The fact that the effect is not immediate or easily measurable does not necessarily mean that it does not exist.

Paragraph changes were occasionally made in the rewrites in response to written comments, but the changes were usually confined to a change of size. The size of paragraphs seemed to be the most common problem. Students started off writing paragraphs that were too long, and then overcompensated by writing paragraphs that were too short. Comments usually drew attention to size rather than to indications of points in the writing where changes should occur. Students were then

expected to make their own decisions about how to solve the problem, which they usually managed very well. Ability to paragraph effectively showed noticeable improvement by the end of the project.

The number of rewrites the students produced increased as time went on, but the number of complaints about the demand for rewrites also increased. Negative comments about having to rewrite the essays were the only complaints the students made during the course of the project. It became clear that what they were doing was not rewriting, but copying the original essays sometimes with corrections, sometimes without, sometimes incorporating new errors, and only very occasionally making worthwhile improvements. It was hoped that the more or less blind copying would develop into something more productive, but for most of the students, most of the time, this did not happen.

The only changes that might be considered more than minor surface changes were changes to the size of paragraphs. Sometimes the paragraphing did not appear to be any better than it had been on the previous attempt, merely changed, but gradually those students who had paragraphing problems improved in this respect. Usually the improvement occurred in successive essays rather than in the rewritten one, and the improvement was not even. Minor surface changes were not always successful because some students copied the original essays including the original errors, appearing not to see the corrections I had marked. Other students copied the original essays incorporating the corrections but making other errors. This seems once again to support the finding that error corrections do not benefit students, although comments on paragraphing did seem to be helpful. It is fairly clear from an examination of the rewritten essays that the rewriting did not result in improved writing and often had the opposite effect.

The practice of asking students to engage in extensive rewriting has followed the insights that writing is generative and recursive, and the fact that 'good' writers make more drastic and more extensive changes to the text than 'poor' writers, who tend to make only minor surface changes. It follows from this that to become a good writer, a student needs to get into the habit of revising and rewriting. The problem here seems to be the difference in perception between the teacher and the

student. If a student perceived a problem with his or her writing, then the writing would be changed. If, on the other hand, it was the teacher who perceived a problem with the writing, the student did not appear able to change it. A student could incorporate the surface changes if there was enough patience with attention to detailed copying, but a major rearrangement of text or a clearer rephrasing of ideas happened rarely.

There seems to be a problem in imagining how others will view a text in the early stages of writing and this difficulty seems to resolve itself only gradually and by means of considerable practice. It seems both from observations in this study, as well as research findings by Allwright (1988) and others, that rewriting does not necessarily benefit students. Most writers remember a time when they did not rewrite their texts after the initial writing. The practice of rewriting, if left to individual writers, seems to develop at an advanced stage of writing, and most advanced writers do not seem to be aware of when exactly the change in practice occurred or of the reasons which prompted the change.

The practice of rewriting is prescribed because the old method of asking students to use the three steps of plan/write/evaluate does not seem to accord with the view that writing is generative and recursive. Careful thought, however, reveals that the steps that do occur, always seem to occur in serial sets. Whenever we write, we enter a state of motivation and then we plan, we write and we evaluate over and over again. What we need to take into account is the awareness that the planning stage cannot be undertaken once only and then forgotten about. There are two kinds of planning - firstly, the planning of the overall text and secondly, the minor planning involved in the gradual realisation of the overall goal and there is the awareness that the evolution of the text may cause the goal either to change, or in some cases, to be abandoned.

The view of the production of writing as a set of three serially ordered steps which recur, not chaotically, but in an orderly fashion at all times is supported by research findings cited by Luria (1973, 1983). Luria's (1973) view of the thinking process is that there has to be a motive, a

restraining of impulsive responses and an investigation of the conditions of the problem. He emphasises that thinking, and by implication, writing, does not end with an answer, but that evaluation always follows on. He cites Anokhin (1963, 1968b in Luria 1973) and Miller, Pribram & Galanter (1960 in Luria 1973) to provide evidence for this view. The gradual change in the recurring operation of plan/write/evaluate occurs as the piece of achieved text grows. The evaluation requirement increases because the new text has to be related not only to the immediately preceding text, but to the whole of the preceding text. This increases the load on STM. It seems that in the early stages of writing development when the writer cannot achieve this, he or she either stops writing altogether or continues without managing an overall view, but with just a local overview. When the writer is still at the stage of local overview, then it does not seem to help for another person to point out a problem with the whole text, because the writer is not yet able to put the matter right. This explains why surface revisions are possible, but not overall revisions for immature writers.

We do not necessarily rewrite, or even edit, a text once the whole of it is finished. Whether we rewrite or edit text that we have already completed seems to depend on four main factors:

1. the length of the text;
2. the kind of text;
3. the amount of involvement the writer has with the text;
4. the amount of time available.

Perhaps the most powerful of these four factors, particularly on the level of difficulty involved in rewriting as opposed to minor corrections, is the length of the text. It seems sensible to suppose that the intention involved in a short text can be more easily managed without rewriting, partly because it can be written at one go. From this point of view it is understandable that beginning writers do not see the need to rewrite since short texts are all that are required of them. Another important factor seems to be the kind of text that is being produced. Narrative writing, for example, usually expects a more familiar audience with plenty of shared reference whose aim in reading is not evaluation. It is easy to appreciate why narrative writing is easier than expository or persuasive writing. The major problems with text coherence occur when the writer progresses to expository and persuasive text. The

third factor, the amount of involvement the writer still has with the text, is most important of all. The writing may be experienced as too difficult and therefore too unpleasant to be worth any further effort. The perception of difficulty seems to depend on length, type of writing and the interest the writer has in what she or he has to say.

The fact that the writer's ongoing involvement with the text affects the amount of rewriting that takes place can be easily acknowledged if we remember the time we take over love letters or apologies, for example. When the content of the text and the relationship with the intended audience is important to us, we take care to write as clearly as possible. Such texts are usually shorter and easier to produce than academic writing, so it is understandable that academic writing may not feel worth the pain of continued involvement. It seems, in any case, that for any type of text, rewriting is likely to be not only useless but counterproductive unless it is the writer, in this case the student, who perceives the need for it.

4.6 Problems and pleasures

There were both problems and pleasures. The three practical problems that were not resolved were:

1. not enough rough paper for preliminary writing;
2. not enough time to write and evaluate;
3. constraints on teaching which were generated by having to conduct the experiment with the control and experimental subjects in the same room.

The first problem was the lack of rough paper. This remained a problem throughout the project. It seemed there was never enough and students continually asked for more. There was a comment in the questionnaires that students had misused the paper, but whether this meant that they had used the rough paper to write letters, to give away, or to roll cigarettes, is not clear. It seems unlikely that the paper was used to roll cigarettes because the scrap was good quality paper and newspaper is usually the preferred kind for rolling tobacco since it tastes better.

The second problem was the lack of time and this was the most frustrating aspect of the whole project for both myself and the students. There was not enough time to get to know the students properly. There was not enough time for conferencing, not enough time for the students to write and not enough time for marking. It is understandable that so little writing takes place in schools, because the time to practise it is not made available, and not only in Papua New Guinea. From the point of view of the students, the lack of time was stressful and frustrating because many felt that they were not doing justice to themselves. They felt they needed longer to produce good writing.

The third problem was the necessity of seeing both groups at the same time. This meant that, apart from general language points that were relevant to both groups, no teaching in the normal sense could take place. The groups were never taught as whole groups and the teaching had to be confined to individual comments made either orally in conferencing sessions, or written as feedback on individual essays. I was prevented from interacting with the groups separately for fear of affecting the other group or mixing the treatments. Group discussion of problems common to the separate groups was prevented and this created a continual frustration. The only advantage it appeared to have was to see whether the students improved with just the comments that were given on individual essays. I felt that it would have been helpful for students to have discussed their problems as writing groups. None of the students commented on the lack of whole group teaching. It could be that I enjoy playing teacher with a whole class more than I realised, but it is more likely that any teacher would find such a situation frustrating. If felt as though a valuable teaching aid was missing, and it was, because the contribution of group interaction was prevented.

The principal pleasure of the writing project followed much the same pattern as the pleasure associated with achieving a single piece of writing. It came at the end. There was an enormous sense of relief that the project had been successfully completed and was over. It became clear from the questionnaire data reported in the next chapter that the students had experienced the writing project as worthwhile, although extremely time-consuming and difficult. My perceptions were similar. The practice of marking a possible total of 134 essays and rewritten essays each weekend on top of a

heavy university workload had made normal living impossible and could not be contemplated as an ongoing activity.

There were other pleasures. There was a great satisfaction in seeing the students' writing gradually improve. Some of the essays were a pleasure to read, and the students' company in classtime was enjoyable despite the frustrations described above. There was a sense of shared achievement because the students knew that their work was improving and this tempered the pain of the effort. Their detailed comments revealed by the questionnaire data will be reported in the next chapter.

CHAPTER 5 - STUDENTS' OBSERVATIONS (QUESTIONNAIRE DATA)

This chapter will report questionnaire data describing student views and observations. The first part will describe general reactions to the writing project, and the second part will report on how students responded to the essay titles given during the project.

5.1 Reactions to the writing project

An anonymous questionnaire was designed to discover student reactions to various types of feedback, as well as perceptions of the 'best' and 'worst' aspects of the writing project, and to elicit assessments of whether or not writing skills had improved, and whether or not the project was enjoyable. Since the questionnaires were anonymous and filled in after the writing project was over, I believe that the answers were honest. It is particularly interesting to see the kind of comments that were made in response to the open questions concerning the 'best' and 'worst' things about the project and to see what students considered worthy of further comment. The response rate was just over two thirds. (See Appendix H for details of the anonymous questionnaire.)

5.1.1 Feedback

Two specific direct questions were asked about feedback:

1. Did you find the essay corrections helpful?

2. Did you find the comments at the end of the essay helpful?

The results are given in Table 11 below.

Table 11: Response to feedback

response rate: 46/68 = 67.6%

	yes	no	don't know
1. Corrections helpful?	46	0	0
2. Comments at end helpful?	46	0	0

There was a unanimous perception that both written corrections and written comments had been helpful. Although this perception may be due in part to the expectation that whatever the teacher

does should be perceived as helpful, it seems unlikely that the student perceptions are wholly attributable to such a cause. For example, the students criticised my belief that rewriting essays would be helpful to them. Written corrections and comments were further singled out for special mention by several students in the further comments section when they thanked me and commented that their writing skills had improved as a direct result of the written comments. Recognition of student efforts both through written comments and grades was specifically mentioned.

For example:

'I liked the good comments and the points I received after a hard work.'

Peer feedback was not mentioned, but peer participation drew comment from one or two students, who wrote that it had been pleasurable to read each other's stories. Conferencing sessions, too, were hardly mentioned, but received favourable comments by the few students who did draw attention to them. Presumably those students, who had probably disliked the conferencing sessions because of repeated nagging about their failure to produce the weekly essay, forbore to comment.

5.1.2 Improved writing skills

Students were asked the following question: Do you think your essay writing has improved? The results are given in Table 12 below.

Table 12: Perception of improved writing skills

response rate: 46/68 = 67.6%			
	yes	no	don't know
Essay writing improved?	44	0	2

Students valued the improvement of their writing skills as one of the most important and valuable aspects of the writing project. Almost all the students who responded to the questionnaire thought that their writing skills had improved. Many commented on specific aspects such as improvement in language skills, paragraphing etc. as detailed below in comments on 'the best thing about the writing project'. One student believed that by the end of the writing project, she could write anything:

'....we all.. had more idea to write stories and poems or anything.'

5.1.3 Enjoyment

The fourth question on the questionnaire asked: Did you enjoy the writing project? Table 13 below shows the results.

Table 13: Enjoyment of the project

response rate: 46/68 = 67.6%			
	yes	no	don't know
Enjoyed project?	27	17	2

About a third of those who responded to the questionnaire stated that they did not enjoy the project. It is clear from comments in other sections that lack of time and general difficulty of writing made the project a painful experience for some students. It is also noteworthy in connection with this perception that these same students still felt that the project had benefited their writing skills. More than half, however, commented that despite the time it took and the pain it caused, it had still been enjoyable. For example:

'...my hands were tired but still the interest kept me to complete the essays.'

5.1.4 Best thing

Table 14 below was compiled by identifying aspects of students' replies to the open question: What was the best thing about the writing project?

Table 14: The best thing about the project

response rate: 46/68 = 67.6%	
	No of comments
Writing (enjoyment & improvement)	12
Story writing (enjoyment & improvement)	10
General English (improvement)	7
Spoken English (improvement)	7
Thinking (improvement)	7
Paragraphing (improvement)	6
Handwriting (improvement)	4
New words (increased knowledge)	4
Reading the stories (enjoyment)	3
Other (improvement)	2
Total number of comments:	62*

*Please note that some students commented on more than one aspect of the project.

The best thing about the project was the sense of improved writing skills. All 46 respondents commented on how pleased they were that their writing skills had improved. For example: *'The best thing about the writting project was that from the beginning my essays were not yet but to the end the essays I wrote were improving...'* Some commented that their English language skills had improved, both spoken and written. Six students mentioned improvement in paragraphing and four thought their vocabulary had increased. Interestingly, several students (7) commented that the best thing about the project was that it had improved their thinking skills. For example:

'...the essay gave us alot of thinking and we had to think deep in our brain before we write the story.'

and

'...it has made me immaginate a lot.'

Four students were even convinced that their handwriting had improved, although I do not remember any evidence of this.

The second best thing about the writing project was that it had been enjoyable. More than half the respondents (25/46) commented that some aspect of the project had been enjoyable. For example: *'The best thing about the writing project is that it is very interesting to write stories about other people or own imaginations.'*

or

'The best thing about the writing project was when we were given funny topics to write about...'

Three students commented that reading the stories was particularly pleasurable. For example:

'I really enjoyed reading my own stories.'

'The best thing was to get to read other students essays or their experience.'

5.1.5 Worst thing

Table 15 below, like Table 14 in the previous section, was compiled by identifying aspects of students' replies to the open question: What was the worst thing about the writing project?

Table 15: The worst thing about the project

response rate: 37/68 = 54.4%	
	No
Time consuming	20
Rewriting	7
Writing about unpleasant emotions	6
Thinking	4
Total number of comments:	37

The worst thing about the project was lack of time. More than half the respondents (20/37) complained that the writing had been difficult because it was time-consuming. There were many comments like the following:

'It took most of both my free and study time during the week.'

There was a sense of frustration at being given an impossible task. The student, who made the following comment obviously wanted very much to be successful, to fulfil the tasks set, but with the best will in the world, could not do so:

'..worst thing..was the non-completion of a handful of essays.'

Some students commented that the time-consuming nature of essay writing had interfered with their other homework. For example:

'..it was sometimes stopping me from studying my note or from doing homework.'

And the following observation is one that every writer can relate to:

'Sometimes it takes me more than ten minutes to complete a sentence.'

One student turned the frustration into a positive perception and came to the conclusion that the aim of the project had been to make him into a speed writer:

'The best thing about the project is to improve the spelling and see how fast we can write.'

Rewriting was the second most frequently mentioned difficulty. About a fifth of the respondents (7/37) identified the burden of having to rewrite the essays as the worst thing about the project. Complaints about rewriting were made not only in comments on the worst aspect of the writing project, but again in the 'Any Further Comments' section. As mentioned in Chapter 4, complaints about rewriting were the only complaints the students made during the course of the project. Rewriting was perceived to be demotivating and unhelpful. Comments like the following were typical:

'The worst thing was after finishing the story, I had to rewrite it in the new page and check through it..'

The key to the problem is probably contained in the above comment which mentions 'finishing the story'. The students perceived their stories to be 'finished', and so at best the rewriting was seen as a copying exercise to incorporate minor surface corrections. There was either no perception of a need to improve the stories or no desire to do so.

Another unpleasant aspect was having to write about topics which evoked unpleasant emotions. Topics which called forth unpleasant emotions were identified by a few students from both groups (two from the control group and four from the experimental group) as the worst thing about the writing project. Frightening experiences were identified as being particularly unpleasant, for example:

'I sometimes think of evil things and get scared.'

In addition, some students identified unpleasant feelings of guilt as the worst aspect. Students from Group 2, the experimental group, wrote that they felt guilty when imagining immoral acts such as robbing a bank, and one or two commented that the act of inventing any stories made them feel guilty. For example:

'...truely speak I felt guilty of making up stories and essays.'

The other difficult aspect of writing that was identified was the requirement to think. For four students from the experimental group 'thinking' was the most painful part of the project and this was frequently coupled with a perception of imagining as a difficult and unpleasant activity. For example:

'It gives me a lot of thinking and having just to many imaginations.'

5.1.6 Any further comments

It was interesting to see that many students had made 'further comments'. Table 16 below was compiled by identifying the main aspects of their comment.

Table 16: Any further comments

response rate: 23/46 = 50%		
		No
	thanks	14
	enjoyment	12
	didn't like rewriting	9
Total number of comments:		35*

*Please note that some students made comments on more than one aspect of the project.

The aspects of the writing project that the students felt needed additional mention were expressions of thanks for improved writing skills, statements that the project had been enjoyable despite the difficulties, and a reiteration that the practice of rewriting had not been helpful.

5.1.7 Summary

The two most criticised aspects of the writing project were lack of time and the requirement to rewrite the essays. It seems that I underestimated the amount of time students needed to write the essays, and had an erroneous view of the value of rewriting at this stage of development. The students did not perceive a need to rewrite. They thought their essays were finished and it seems that their perception is what mattered since the rewriting rarely seemed to achieve any improvement.

Most students appreciated the writing practice despite the frustration experienced because of lack of time to write, and despite a perception that writing essays was difficult. Many students enjoyed writing the stories and almost all felt that their writing skills had improved. There was a unanimous perception that teacher feedback had been helpful. Many made perceptive observations about the mental processes that had occurred during the writing. In particular they commented on their feelings when writing various types of essay and the fact that they had learned through writing. The fact that they commented on writing as a thinking and learning process was surprising since the idea that writing promotes thinking and learning had never been mentioned. I found their awareness and analysis of mental processes impressive.

5.2 Response to the treatment titles

This section will report first of all on how students responded to the essay titles given during the writing project, and will then discuss the mixing of writing types that the data revealed. A questionnaire for each group (see Appendix F) was designed in response to noticing that one or two students in the control group had written ISN essays instead of PHN. The questionnaire was given at the end of the writing project and had three aims:

1. to check for mixing of writing types;
2. to find out what the students' response had been to specific titles;
3. to see how the students evaluated the tasks of their own group compared to the tasks of the other one.

The most important of these aims was the first one: to find out whether the groups had written PHN or ISN, since practice in the different types had been prescribed by the experiment in order to determine whether one kind of practice was more beneficial than the other. From the point of view of the success of the experiment, it was disappointing to discover that the writing types had not always been the ones that had been intended. From the point of view of finding out about what actually happened during the development of student writing skills, the data from the questionnaires were illuminating.

5.2.1 Mixing of writing types

One of the primary aims of the experiment was to compare the effect of practice in one kind of writing with the effect of practice in another. The control group was intended to receive practice in PHN, while the experimental group would receive practice in ISN. It was acknowledged that a third type of narrative existed, OPN (see Chapters 1 and 3), but I had intended to exclude this kind of writing practice from the experiment. I had believed that the type of writing practice could be controlled by choosing the essay titles carefully, by asking experienced Papua New Guinea teachers to check the appropriateness of titles, and by giving instructions to the students about the kind of writing required. The data from the student questionnaire showed that this belief had been an illusion.

5.2.1.1 Control group

Students were asked in the questionnaire to quantify the amount of personal experience recounted in each essay. The following instruction was given:

Check each essay in your exercise books and decide whether it was:

T true - a retelling of your own experience

PT partially true - a retelling of your own experience with some imagined bits

SS somebody else's story - a story that you had heard or read or seen on television

I imagined - an imagined experience (it did not actually happen)

The response of the control group on mixing of writing types is summarised in Table 17 below.

Table 17: Mixing of writing types - control group

	T	PT	SS	I	response rate	
					no	%
1 Escape from danger	6	2	3	14	25	73.5
2 My life story	21	3	0	2	26	76.5
3 Worst thing I ever did	10	3	3	9	25	73.5
4 First time I watched television	5	4	3	14	26	76.5
5 Exciting journey	8	5	1	11	25	73.5
6 Child minding	8	4	4	10	26	76.5
7 Mysterious place	3	5	5	13	26	76.5
8 A student's day	12	3	1	10	26	76.5
9 A funny thing	10	5	1	7	23	67.6
10 Friend in need	5	8	2	9	24	70.6
11 Best letter	9	6	0	11	26	76.5
12 Storm	10	3	0	13	26	76.5
13 Bad deed	8	3	1	13	25	73.5
14 My revenge	5	5	3	13	26	76.5
15 Exciting ride	13	3	1	9	26	76.5
16 Fishy story	11	6	0	8	25	73.5
17 Handicapped friend	3	5	0	16	24	70.6
18 Frightening experience	7	4	3	11	25	73.5
19 Hurt in an accident	7	4	3	11	25	73.5
20 Memorable shopping trip	12	4	1	8	25	73.5
TOTALS	173	86	35	212	506	74.4
Proportions	34.2%	17%	6.9%	41.9%		

It seems that the control group's treatment titles elicited the desired writing type in only a third of reported cases. The titles elicited more invented experience of the ISN kind (41.9%) than PHN (34.2%). It seems that the third type of narrative, OPN, was not excluded either (6.9%). Since the students appeared to have a clear concept of the difference between PHN and ISN, it is unlikely that the problem was due to a lack of understanding of what I was asking for. There seem to be four possible reasons:

1. a perception that ISN was valued more highly than PHN;
2. failure to find a match between title and personal experience;
3. failure to find a match between personal experience and the perceived expectation of the teacher;
4. reluctance to write about personal experience when the title elicited unpleasant memories.

Reason 1: A perception that ISN was valued more highly than PHN appears to be the least likely reason for its inclusion. I was aware from the outset of the need to treat both groups with equal enthusiasm and this turned out to be easy since students in both groups produced interesting essays

that I enjoyed reading. It is clear, too, from the students' responses to the part of the questionnaire which asked about their attitude to the tasks of the other group that such a perception was not conscious. Despite such evidence to the contrary, however, the explanation remains a possibility since there is no way of 'proving' that this was not the case.

Reason 2: It seems that there were cases where there may have been no corresponding personal experience to match the title. An example of a title where this seems to have happened was the 'Handicapped Friend' title no 17. I had been assured by PNG colleagues that all students would have had experience of living closely with some handicapped people in their villages, but the students' response to this title suggests otherwise. Student response suggested either that there had not been any handicapped people in the village, or that the students had not considered them friends.

Reason 3: It became clear that there were some occasions when the student experience which matched the title was discarded due to a perception that the experience would not suit my expectation. There are at least two examples where this seems to have happened: Title No 1 'Escape from danger' and Title No 4 'First time I watched television'. From discussions with comparable classes of PNG high school students, it seems that all children have had escapes from dangerous experiences, such as escapes from being killed by snakes, by spiders, by drowning etc. and PNG colleagues confirmed this view. In the case of the Laloki High School students in the control group, it seems more likely that the students felt that such experiences were not what was wanted rather than that they had not had such experiences. Similarly the request for an account of the first time they had watched television produced largely invented accounts, despite the fact that the school possessed a television expressly for students so they would all have experienced a first-time TV viewing.

Reason 4: It seems that sometimes a title called forth an unpleasant memory which the student did not wish to recount. Title No 18 'A frightening experience' is a good example of where this might have occurred. Although we can assume that everyone will have had a frightening experience at

some point in their lives, most of the students chose to invent experience for this essay, rather than to recount their own. Comments on other parts of the questionnaire confirmed that some students did not like to write about unpleasant experiences.

In summary, it is clear that the reason for writing about something other than one's own personal experience can be motivated from within or from without. It can come from within because of an inability to write about the topic given or because of a reluctance to write about the topic. It can come from without as a response to perceived audience influence in order to present the audience with something more suitable than the actual personal experience which is available, something more impressive, more entertaining or more respectable perhaps. It seems that although some titles worked better than others to produce an account of personal experience, no title was foolproof. It does not seem possible to anticipate all the individual differences of experience and attitudes to those experiences. Students took the obvious way out when the required personal experience was difficult to recount: they invented something.

5.2.1.2 Experimental group

Students were asked in the questionnaire to quantify the amount of personal experience recounted in each essay. The following instruction was given:

Check each essay in your exercise books and decide whether it was:

I totally imagined	- an imagined story invented by yourself
PI partially imagined	- a mixture of real and imagined experience
SS somebody else's story	- a story that you had heard or read or seen on television
T true	- your own experience

The results are given in Table 18 below.

Table 18: Mixing of writing types - experimental group

	I	PI	SS	T	response rate	
					no	%
1 Escape from the sea	20	3	1	0	24	70.6
2 Life story of beautiful person	20	2	1	0	23	67.6
3 Day I robbed the bank	10	1	3	1	15	44.1
4 First television appearance	21	3	0	3	27	79.4
5 Trip to Midwinkle	20	2	0	1	23	67.6
6 Looking after Colin	9	7	0	7	23	67.6
7 House of happiness	18	3	4	1	26	76.5
8 Day in Prime Minister's life	22	2	1	0	25	73.5
9 Teacher who made us laugh	12	6	2	3	23	67.6
10 Letter that changed my life	18	6	0	1	25	73.5
11 Ada, the helpful spirit	20	3	3	0	26	76.5
12 Storm that destroyed PNG	20	3	1	0	24	70.6
13 Wicked woman	14	5	4	1	24	70.6
14 Tapoi's revenge	19	3	4	1	27	79.4
15 First driving lesson	18	1	1	2	22	64.7
16 Mermaid	19	1	3	0	23	67.6
17 Blind	19	0	2	1	22	64.7
18 Night bird	9	1	2	1	13	38.2
19 Buried alive	21	1	1	0	23	67.6
20 Million kina shopping spree	6	0	2	0	8	23.5
TOTALS	335	53	35	24 (447)		
Proportions	74.9%	11.9%	7.8%	5.4%		

In contrast to the PHN titles that elicited the desired writing type in only a third of the cases, the ISN prompts were effective in almost three quarters of reported cases. It seems to be much easier to elicit invented writing than to be sure of eliciting 'personal experience' writing. Does this mean that students (and writers in general) are so used to 'adding a little embroidery' that they usually cannot manage to write without creating an improvement on original experience?

The writing type that might have been expected to give the greatest trouble in the experimental group's production of ISN was the category of 'other people's stories'. According to the questionnaire data, this does not seem to have been the case as OPN represented only 7.8% of reported cases. This may not, of course, present a totally accurate picture of what happened. Plagiarism is a Western concept, which is not recognised as 'bad' in traditional PNG culture. The students in PNG are generally aware that Westerners consider 'copying' as bad, so they usually smile sweetly and deny having 'copied', even when accounts are identical. There were no cases in the

writing project where students produced the same essays as each other, but it cannot be ruled out that they did not on occasion use the stories from films or from other people and present them as their own. The proportion of essays where students claimed to have written PHN is smaller still than the OPN category, just over 5%. I doubt, too, that this proportion is entirely accurate, since almost all the stories students produced seemed unlikely to have happened. The claim to have produced PHN stories may stem from a common PNG belief that fantastical experiences can be real. The stories may have become so real to the students, that they thought the experiences had actually happened.

Overall, the ISN treatment titles worked well to elicit the desired writing type. The problem for the experiment was that there was no longer a clearly contrasting different writing type practice with which to compare it, since the PHN titles had elicited more accounts of invented experience than of actual experience. I was left with the knowledge that the experimental group had had *more* practice in imagined story writing than the control group, but that a clear division between the two groups had broken down.

5.2.2 Response to specific titles

5.2.2.1 Control group

In addition to identifying the kind of experience used for each essay, students were invited to make comments on each one. The questionnaire also asked them to state their favourite and least favourite essay and give reasons. A summary of the results for the control group appears in Table 19 below.

Table 19: Response to PHN titles

Favourite Essay response:24/34 =70.6%		Worst Essay response:22/34 =64.7%	
Fishy story	6	Fright	3
Life story	5	Child minding	3
Exciting ride	3	Escape from danger	3
Funny thing	2	Funny thing	2
Letter that changed life	2	Storm	2
Friend in need	2	Handicapped friend	1
Escape from danger	1	Mysterious place	1
Mysterious place	1	First time for TV	1
Shopping trip	1	Bad deed	1
Bad deed	1	Student's day	1
		Worst thing	1
		Revenge	1
		Fishy story	1
TOTALS	24		22
Reasons		Reasons:	
enjoyed experience	16	disliked experience	7
because it was true	5	because imagined	6
easy to write	3	boring	5
liked reading story	1	time consuming	3
learned from experience	2	difficult	2
		low mark	1
(note: some students gave more than one reason)			

There was a wide range of response to specific titles. A title that one student might find to be the worst essay title might be experienced by another student as the best. This emphasises how difficult it is to predict the response to a particular title because the writer's range of experience and attitude to experience is an individual affair and capable of wide variety. The reasons for their response, however, were amenable to generalisation. The main reason given for a title being either most liked or most disliked was that the experience to be related was either enjoyable or unpleasant. For example, reasons for favourite essays:

'Because I really enjoyed myself at that time.' (Fishy story)

'Because it was an exiting trip I ever had in my life out at sea.' (Fishy story)

'I really love this story because it brings back memories..' (Life story)

Reasons for worst essays:

'Because it is the scaring place where ...people go and get lost...'(Mysterious place)

'Because when I read it, I get scared in the night and I never want to go to bed/sleep.' (Mysterious place)

'Because it makes me angry....' (Revenge)

'I didn't enjoy the story because I regret what I have done.' (Worst thing)

'Next time I won't go out like that. Better store it in my head.' (Friend in need)

The role of emotion is important to note. The conclusion is simple: if an experience was pleasant, the story was perceived as easier to write than if the experience was unpleasant. What is interesting in relation to this finding is the consideration of whether a writer has a need to write about pleasant experiences. There is a common belief that painful experience produces the best writing. At this level of writing development, there was no evidence for such a view.

The second most common reason for either liking or disliking a title was related to Papua New Guinea cultural values. Students explained in the questionnaire that the factual nature of a story made it desirable, while the imaginary nature of a story tended to make it undesirable. Such observations were not totally unexpected, given the Papua New Guinea attitude towards invented stories discussed in Chapter 1. Students had, however, not made such comments during the project. When the students in this group made such comments after the project, they presented the view as a self-evident fact. They did not explain why the factual nature of the stories should make them automatically more desirable. Their comments on this issue present a contradiction to the fact that they frequently chose to include invention even though instructed not to do so. The following are examples of student comments:

'I like it because there wer facts written down.' (Fishy story)

'Because I told truth and everything was there... I tried to think of what was next.'

'Because it was true indeed, it made me really think back and laugh.' (Funny thing)

'it was only a imagination one..' (Escape from danger)

'I don't like it because it was someone's experience I wrote about' (The shopping trip)

'because I told lies...' (Exciting trip)

In addition to disliking imagined stories because of their lack of truth, there was also a fear of imagined experience: '*Because I have never experience on of these things till now and I don't want it to happen to me to experience.*' (Mysterious place) However, not everyone in the control group saw imagination as unpleasant or dangerous. Another student who commented on the same essay title commented that he had used his imagination to remove himself from a frightening experience: *'I decided to place myself in my friend's story instead because the first time I visited that place at night I did have somewhat similar reactions towards that place so I decided to step into his shoes for a while....'* And yet another student commenting on the same title had enjoyed exercising his imagination for reasons of power: *'I really like this imaginary story because some people mistake it as a true story.'* (Mysterious place)

In connection with the value attached to the truth of a story, it is interesting to note that the control group chose 'A fishy story' and 'My life story' as their favourite titles. Both these titles seem to have been particularly enjoyed because they were successful in eliciting personal experience. Most of the students were from the Papuan coast, so the story about a fishing trip appealed to them. A typical comment on why this title was the favourite: *'Because I was involved in fishing and I wrote it easy.'* And a typical comment on the 'Life story' essay was: *'Because all the things I wrote in that essay were all true or facts about my life.'*

Another reason given for appreciating a particular essay was that the student had learned from the experience or because they felt that the telling of the story had been beneficial:

'...I wanted to portray my innermost emotions.' (Friend in need)

'....made me realise nothing comes on a golden plate.' (Letter)

'It was somewhat shamefully for me myself but much easier to express in ink than in words so I healed my experience....' (First time I watched television)

Finally, students obviously liked stories that were easy to write and fun to read:

'Because I wrote a funny story which made me laugh when I read it after finish.' (Funny thing)

'Because it took me less time to complete and words were coming out in my mind very quickly.'

(Fishy story)

In summary, the most powerful reason for a positive response to a PHN title seems to have been because the title elicited a pleasant memory and the second most important reason was 'because the story was true'. These perceptions were mirrored exactly in the reasons given for not liking a particular title i.e. either the writing was about an unpleasant experience or it was 'not true'. It is interesting to consider these comments on the pleasure of writing 'true' stories together with the fact that two thirds of the control group, who were supposed to be writing about personal experience, chose to include invention in their essays.

5.2.2.2 Experimental group

A summary of the response to ISN titles is given in Table 20 below.

Table 20: Response to ISN titles

Favourite Essay		Worst Essay	
response: 22/34 =64.7%		response: 18/34 =52.9%	
Letter	3	Driving lesson	3
Driving lesson	3	Buried alive	3
Beautiful person	2	Robbing bank	3
Escape from sea	2	Blind	2
Blind	2	Beautiful person	2
Minding Colin	2	Shopping spree	2
Shopping spree	2	Minding Colin	1
Night bird	1	House of happiness	1
Prime Minister	1	Wicked woman	1
Funny teacher	1	Storm	1
Mermaid	1		
Robbing bank	1		
Storm	1		
TV appearance	1		
TOTAL	22		19
reasons:		reasons:	
enjoyed experience	14	disliked experience	10
good mark	3	hard to imagine	3
liked reading story	3	low marks	3
easy to write	2	boring	2
thought provoking	2	time consuming	1

As with the control group, the students in the ISN group, identified the main reasons for liking or disliking a particular essay title as the enjoyment or unpleasantness that the experience evoked. In this case most of the experience referred to was imaginary. The comments made it clear that imagined experience can be very powerful. Typical comments on good experiences while writing imagined stories were:

'When I was reading this essay I was laughing away to myself.' (Funny teacher)

'They were my favourite essays because they were funny and like they were the imaginations of my future..' ('Funny teacher' & 'TV appearance')

'I had high feelings when I wrote it...' (Most beautiful person)

'I tried it out...' (Looking after Colin)

'I laughed as I wrote it..' (Funny teacher)

'I like it because I was the only won who was alive...' (Storm that destroyed PNG)

In contrast, if the experience the students were writing about was perceived as unpleasant, they did not like writing the essay. For example:

'I don't like it because I don't want to laugh in my lesson time...' (Funny teacher)

'I didn't like it becaues Iam getting afrid of spirit.' (Ada, the helpful spirit)

'I was sad when I wrote the story and when I just imagine it I feel like crying when some people lost thier life.' (Storm that destroyed PNG)

'I didn't like it because I'm not a blind person after all.' (Blind)

'Truely I didn't feel like writting it because what if it was true but anyway it was a must but otherwise it was good because it made me laugh...' (Buried alive)

'I don't like it because while I am alive...they had buried me.' (Buried alive)

An observation that was made about writing ISN, that was not made in relation to PHN, was that it caused deep thinking. The need for thinking when writing imagined stories was commented on from both a positive and a negative point of view. For example:

'I really like it because ....it made me think further and deep about things..' (First television appearance)

'...it was bit difficult for me to write about a place I had never heard of.' (Trip to Midwinkle)

'It was a bit hard for me to write this story, it made me think alot.' (Looking after Colin)

In addition to comments about writing the essays, some students commented on the pleasure of reading them: *'It was very interesting to me and I felt like reading the story the whole day.'* (Storm that destroyed PNG). There were comments, too, that echoed the remarks the students in the control group had made about guilt and fear being associated with imagination, e.g. *'I did not really like this story because it made me feel guilty about stealing.'* (Robbing the bank)

For some students, it was as though the act of imagining had been a means of discovery and coping, e.g. *'I like it even though the thought of being buried alive made me scared.'* (Buried Alive)
It is necessary to remember, however, that the ISN essays contained some true experience as well as some input from other people's stories in addition to the experience that had been invented.

In summary, the ISN group had a slightly wider range of favourite essay titles than the PHN group. The reasons they gave for liking or disliking essay titles were broadly similar to those given by the control group students, despite the fact that they were usually referring to a different kind of writing practice. The main difference in the reactions to specific essay titles was that the experimental group did not make negative comments about having to invent stories. It seems that either the students got used to inventing stories and changed their attitudes towards the practice, or, less likely, had had different attitudes from the control group students all along. A third possibility is that the students did not like inventing stories, but did not say so in the questionnaires, but this possibility does not seem likely in view of the honesty of the student comments in general.

5.2.3 Observations on the other group's tasks

5.2.3.1 Control group

Only 3 out of 21 who responded said that they would have preferred to be in the group which wrote imagined stories. All three explained that they would have preferred the other group because '*there*

were some nice stories to write about.' The majority were happy with the group they were in and gave as their main reason the fact that they preferred their own essay titles. Several students (6/21) commented that the other group's essays were more difficult. This was an interesting comment in view of the fact that the groups had been told that the level of work would be the same for each group. It is also worth noting because it was a comment made only by the PHN group about the ISN group, and not vice versa. Typical comments were: *'..because their essays are bit difficult and won' t suit me.' or '..because it would be twice worser if I rather have been in group two and maybe it...might have difficult essays to write about...'* Some commented that they did not like the other group's essays because they required imagination: *'because the essay ....were mostly imaginary ones'*

Many students, however, gave positive reasons for their preference such as *'Because I liked been in Group 1 and like the essay topics and were bit easy too.'* or *'Because the story headings I found in Group 1 were really interesting for me to write about..'* or simply *'...I enjoyed the essays...'*

5.2.3.2 Experimental group

Of the experimental group, only 2 out of 28 who responded would rather have been in the PHN group and both students gave as their reason that the imagined story narratives were too difficult for them. The rest of the students made positive comments and most said they preferred the ISN group because they found the essays more interesting.

5.2.4 Summary and discussion

The two most noteworthy observations concerned the mixing of writing types and the emotional effect of the essay prompts. It is interesting that far more invented experience went into the personal history narratives, than the other way round. It was clear, too, that 'other people's stories' had been used to some extent. The observations made clear that it is not possible to ensure that students produce a particular writing type merely by instructing them to do so. The honesty of comments expressed in questionnaire date can always be questioned, but my own belief is that the students were telling what they believed to be true.

Both groups wrote better and more willingly about pleasant experiences than about unpleasant ones, whether these experiences were real or imagined. This finding is supported by neuropsychological evidence cited by Damasio (1994) that happy cognitive states produce better reasoning than sad cognitive states. He emphasises the difference in cognitive process between a happy state and a sad state:

> '... the cognitive mode which accompanies a feeling of elation permits the rapid generation of multiple images such that the associative process is richer and associations are made to a larger variety of cues available in the images under scrutiny ... This cognitive mode is accompanied by an enhancement of motor efficiency and even disinhibition, as well as an increase in appetite and exploratory behaviours ... By contrast the cognitive mode which accompanies sadness is characterised by slowness of image evocation, poor association in response to fewer clues, narrower and less efficient inferences, over concentration on the same images, usually those which maintain the negative emotional response. This cognitive state is accompanied by motor inhibition and in general by a reduction in appetite and exploratory behaviors.' (1994:163-4)

The perception that the personal experience that was asked for was unpleasant seemed to be one of the strongest reasons for including invented experience in PHN. Other reasons for including invention seemed to involve a feeling that the audience would prefer something other than the personal experience that was available. ISN was easier to elicit, but, once again, it was not possible to ensure the students invented their own stories, nor was it possible to control for the amount of personal experience that was included in the essays. It is important to note that the findings do not imply that the two narrative types compared in the study do not exist. The implication is rather that they may never be totally discrete. At the level of development investigated in the study they were heavily mixed, but, as mentioned above, not to the same degree. The findings emphasise that all three narrative writing types, which were identified and described in Chapters 1 and 3, do exist and that writers are both aware of, react to, and have a use for, different narrative writing types. The student observations on how they tackled the essays are probably the most valuable findings of the experiment. This is because the information revealed that the control of different kinds of writing practice had not been effective and that writing types had been mixed, and because it emphasised that the emotional response of the student had a profound effect on writing and learning.

PART 3: RESULTS AND DISCUSSION

CHAPTER 6 - MEASUREMENT ISSUES

There are three measurement issues that deserve description and discussion so that information can be borne in mind while considering the results of the experiment. The three issues are: 1) the implication of the mixing of writing types during the treatment period on what was measured in the pretests and the posttests, 2) the effect of the essay test prompts, and 3) inter-rater reliability.

6.1 The mixing of writing types

6.1.1 Narrative types

It is clear from the questionnaire data reported in the previous chapter that there was some mixing of narrative types during the treatment for the experiment. This knowledge has two effects on the consideration of the results -

1. The differences between groups by the time of the posttests may be attributable to more or less of a particular type of narrative writing practice rather than solely to one particular kind of practice as originally intended. In other words, the control group had more personal history narrative practice and the experimental group had more imagined story practice. The differences in average group performances between pre and posttests will measure the difference in writing practice focus.
2. It can be assumed that there was some mixing of writing types during the tests as well as during the treatment time.

Given that it is probably not possible to keep writing types separate, at least at the level of writing development investigated in this research, the persuasive writing scripts were also scrutinised. It was found that there had been some mixing of writing types within them too.

6.1.2 Persuasive writing

Some of the persuasive writing scripts contained evidence of narrative insertions. Consider for example the following extract taken from subject 59 on the pretest who was writing the essay titled 'Alcohol should be banned in PNG':

'Secondly drinking alchol will also cause families to tear apart as well and also cause families to face financial problems Eg: The man uses up all the money to buy alchol and drinks with his friends and enjoyes himself, after drinking he comes home and beats up the wife and chilren, The wife then packs up and leaves him while he is heavily drunk - when the weekend is over, he goes to the kitchen and finds breakfast - Nothing is found, he checks his bag not even a coin is inside - He doesn't go to work because of financial problem he faces, at the end of the week, He recieves a letter from he's boss saying - DON'T BOTHER COMING BACK.'

It is clear that the example which occupies most of the above paragraph is a piece of narrative writing. It is true that an attempt has been made to achieve generalisation by changing the verb tense from past to present, but it is nevertheless an example of an extended piece of narrative which has been used in a persuasive essay.

6.1.3 Implications of mixing

It seems that writing types were never totally discrete and as demonstrated above, it was not only the narrative types that displayed evidence of spontaneous insertions of each other, including the third 'banned' type, OPN, but the persuasive essays, too, showed evidence of mixing. It seems likely that immature writers are more likely to mix writing types than experienced writers. In research that seeks to discover how transitions occur as students tackle more difficult types of writing, the fact that writing types have been mixed is important to note.

The awareness that writing types did not remain discrete at this level of writing must be kept in mind when considering the results of two of the experiment's questions: the relationship between writing types, and the attempt to see whether practice in imagined story narrative enabled the transition to persuasive writing more than practice in personal history narrative.

Interpreting the relationship between writing types

Since the writing types that are being compared have been produced by immature writers and are not discrete types, a result that shows no difference between types could mean one of two things. It could mean *either* that there was indeed no difference between the types of writing, *or* that there was such heavy mixing of writing types in the scripts under scrutiny that a difference which did exist between the types could not be seen at the level of writing maturity used in the experiment. A result which *does* show a difference between writing types will reveal that there are indeed significant differences between the writing types which could have been expected to have been even more marked if the writing types had been less mixed.

Interpreting the effect of practice in imagined story narrative

Since there was no pure and exclusive practice in either imagined story narrative, nor in personal history narrative, what is being compared between the groups are the relative benefits of having had a main focus for writing practice on either invention or on description of personal experience. A result that shows a difference between the writing development of the two groups of students will reveal that there are indeed significant differences that are probably attributable to a focus on one kind of narrative practice as opposed to another.

Having established what exactly is being measured and compared in the experiment, the next section will describe some differences in the effects of the test prompts.

6.2 Effect of test prompts

There are two major sources of variability concerning the effect of essay prompts:

1. the personal response of the writer to the content required by the prompt, and
2. the inbuilt cognitive requirement of the essay prompt specified by the teacher or test setter.

The ways in which these sources of variability were taken account of by the test prompt design will be discussed. This will be followed by a report and discussion of the results of the topic effect on performance in the pretests and the posttests for each of the three writing types.

6.2.1 Personal response

The personal response of the writer to the content of the essay prompt, the first source of variability, was discussed in the previous chapter. It was made clear that it is not possible to choose a title which will be sure to suit all writers because of the wide variety of personal experience and attitude to that experience. The writer's response can also determine to some extent the kind of cognitive process required to produce the piece of writing. For example, it was shown in the previous chapter that a writer can choose to invent experience rather than recount actual experience if this seems to be a better option. This could arise either for reasons of personal writing comfort, or from a consideration of what the reader might prefer. This first source of variability is outside the tester's control and it is recognition of this kind of variability that has caused some researchers (Purves 1992, Lumley & McNamara 1995) to advocate that multiple pieces of writing be required from each candidate if testing is to be fair.

6.2.2 Cognitive requirement of prompt

The second source of variability concerns the difference between the kind and quantity of cognitive process that is implied by the essay question. The second source of variability should be amenable to control by the tester. Horowitz (1989) investigated the function and form of 284 essay examination prompts, and identified many differences between prompts. He advocated teaching students, especially non-native speaker students, to understand the requirements of prompts. He did not, however, report on effects of the differences in prompts.

It seems sensible to believe that no two essay writing tasks make identical requirements on a writer, but in an experiment which seeks to quantify improvement over time on a certain type of essay

writing, the tasks on pretest and posttest essays were intended to be as similar as possible. The titles for the pretests and the posttests were drawn up with regard to:

1. fulfilling the aims of eliciting the particular writing types as defined in the research design;
2. providing topics that would be equally familiar to all students
3. controlling for demands of audience
4. controlling the amount of choice available

A fifth consideration should have been an attempt to control for the emotional response of the writer, but the prompt design did not do this. I was aware of the importance of the *strength* of emotional response to a topic, but made no attempt to control for the *kind* of general emotional response a prompt might produce. I did not know it was important to be concerned about whether the topic would be pleasant to write about, or unpleasant, yet this turned out to be important for narrative writing.

A consideration of the effect of essay prompts shows that the first aim, i.e. the intention to control the type of writing elicited, was not totally successful. In view of the feedback on response to the treatment essays, it seems that it is not possible totally to control the type of writing through title choice with writers at this level of maturity. Stringent efforts were made to control the essay topics for levels of familiarity as described in Chapter 3.5.1. This was obviously easier to do for personal history narrative and persuasive writing, than for imagined story writing. Invented experience implies by its nature a lack of familiarity and the degree of novelty is a matter for individual choice.

The demands of audience obviously varied according to writing type, but the audience demands for prompts for the same writing type should have been kept as similar as possible. Audience specifications for PHN and ISN were not made explicit, since they were considered to depend on the writing function. This was the same for both pretests and posttests. The persuasive writing prompts, however, varied between pretests and posttests in that audience for the pretests was implicit, but for the posttests was made explicit. This is discussed in 6.2.3.3 below.

Students were not allowed a choice of title because three titles were devised for each writing type in an attempt for control for topic effect. This meant that the titles had to be allocated randomly and as equally as possible. The unintentional effect of this was to force some students to write about topics they found unpleasant and thereby reduce group performance on such titles. The type of emotion likely to be aroused by the topic turned out to be important, since any differences in performance, though not statistically significant, appeared to be due to differences in the kind of emotional response the topics had evoked. It may be, however, that statistical tests of average performance are too blunt an instrument to reveal some differences that may be important. It is apparent from the comments made by the students, reported in the previous chapter, that a single title was capable of calling forth a wide variety of writer response and individual differences are not usually revealed by statistical tests. It is clear, too, that some titles were generally perceived as either pleasant or unpleasant to write about and that this affected performance noticeably if not significantly. A further consideration is that testing for reliability across three titles in the pretest and the posttest does not reveal differences in essay prompt requirements between pretests and posttests, since such differences in performances are assumed to be the result of the treatment. It is important then to consider not only the results of the statistical tests which show differences in relative performance between tasks, but also to consider individually the prompts used in the test and to speculate on the probable effect of those prompts according to the criteria listed above.

6.2.3 Effect of test prompts on performance

Analysis of Variance was carried out to test for similarity of performance on the pretest and posttest set of three titles for each writing type (significance level $p<0.05$).

6.2.3.1 PHN titles

Aim of titles: to elicit narrative about events that have been experienced personally.

The results are shown in Table 21 (pretests) and Table 22 (posttests) below.

Table 21: Performance similarity on PHN pretest titles

	F	p			
PHN pretest titles	0.31	0.733			
			n	mean /15	stdev
House building celebration			22	9.273	2.142
Harvest celebration			23	9.217	1.999
Bride Price celebration			23	8.783	2.679

There was no significant difference between performance on the pretest titles.

Table 22: Performance similarity on PHN posttest titles

	F	p			
PHN posttest titles	1.47	0.236			
			n	mean /15	stdev
First schoolfriend			23	11.435	1.854
Best present			22	11.000	1.746
Worst punishment			23	10.609	1.234

There was no significant difference in performance on the posttest titles, although the average performances show that 'The first schoolfriend' produced the best results, followed by 'Best present', while the 'Worst punishment' essay was written about least well. This order supports student comments that they preferred to write about pleasant experiences rather than unpleasant ones.

It is clear from student reports on the treatment titles, as already discussed, that it is not possible to be certain that instances of ISN or OPN were not included in some of the personal history narratives. The fact that the titles were chosen to ensure a match between the writing topic and the student's experience did not necessarily ensure that the student used only his or her own experience to write the essay. The 'Best present' title seems to have caused problems. Some students appear to have understood audience expectation to be that the present should be some large material gift, and this caused them to invent rather than recounting actual experience. For example, two students reported receiving a car, which were obvious flights of fancy, while several others reported receiving radio cassette players, which seem unlikely to have been real presents.

I believed that all the essay prompts provided familiar topics to write about, in the sense that all the students would have had experience in all the topic areas chosen for the essays (see Chapter 3.5.1) Experience, however, was bound to be varied. Students may have experienced one type of occasion more than others, or have may have experienced some occasions more intensely and memorably than others. On reflection, it became clear that the titles themselves implied differences I had not thought about at the time of the initial design. For example, one would expect that there would be only one experience of a first schoolfriend to recount, whereas there might be several memorable bride price celebrations to choose between. These titles appeared in different sets so it is not possible to be sure whether they caused differences in performance or not.

6.2.3.2 ISN titles

Aim of titles: to elicit narrative about events that have been invented by the writer

The results are shown in Table 23 (pretests) and Table 24 (posttests) below.

Table 23: Performance similarity on ISN pretest titles

	F	p					
ISN pretest titles	1.73	0.185					
		n		mean /15		SD	
Day in life of bird			23		9.522		1.702
Day in life of fish		22		9.000		2.430	
Day in life of pig		23		8.435		1.754	

There was no significant difference in performance on the pretest titles, although the ordering of performance, like the PHN titles, supported student comments that pleasant experiences were easier to write about than unpleasant ones. The students did not like imagining themselves as pigs. Pigs, unlike birds and fishes, were perceived as dirty, ignorant creatures. For example: *'I am an animal which is called a pig, as you know I'm the dirtiest and unhealthy animal in the world. I have ugly looking face with two legs and hands which I use to walk on.....I'm feeding on rubbish which are lying on the ground ...food scrapes, rotten leaves,..and food grains which people have thrown them*

away.' (Subject 51) Compare, for example: *'I am a bird with long feathers, with very beautified, different colours like the rainbow. I feel proud of myself too.'* (Subject 4)

Table 24: Performance similarity on ISN posttest titles

	F	p			
ISN posttest titles	0.72	0.489			
			n	mean /15	SD
Secret friend (talking dog)			23	11.043	1.331
Unusual present (silver ball with 9 knobs)			23	10.609	2.126
Royal punishment (given to untrustworthy servant)			22	10.455	1.565

There were no significant differences in performance, but once again the ordering supports student comments on emotional response. The most popular title, judging by the level of performance, was the requirement to imagine a secret friend, while the least popular was the requirement to imagine punishing someone.

It is fairly easy to ensure fictional, as opposed to factual, narratives with some choices of essay title, but the danger here is that other people's stories may be used as well. It is clear from the students' questionnaire comments that although titles designed to elicit ISN usually do so, actual experience may also be included. According to student comment on the treatment titles, they very rarely used OPN but this may not be true.

Both pretest and posttest titles appeared to fulfil the requirement to elicit invented experience, as far as such a function can be ensured by the essay prompt and judged by the resulting content, but on closer examination there seemed to be other differences between the prompts. For example, the three pretest prompts required the writer not to be human i.e. to be a bird, fish or pig, while the three posttest essay prompts required a story about the invented experiences of a human being. It could be argued that it should be easier to imagine being a queen or king for example, than to imagine being an animal, but the dimension of familiarity complicates matters. For PNG students, the lives of pigs,

fishes and birds are, at least superficially, familiar to the students, whereas the lives of royalty are not. In terms of the amount of invention needed it might be easier to quantify essay prompts in terms of the degree of familiarity of topic, but it is not easy to assess relative degrees of difficulty. Compare for instance, the problem of writing familiar events from an imagined animal persona, compared with the difficulty of imagining human reactions to a talking dog, or the requirement to explain the purpose of a silver ball with nine knobs.

It is not possible to make performance comparisons between pretest and posttest title effects because of the difference made by the treatment, but it is worth noting that there was no significant difference in performance between the 'Royal Punishment' title and the 'Unusual Present' title, which were both posttest prompts. Yet the 'Royal Punishment' essay required students to imagine they were someone with high status, in effect a different person, while the 'Unusual Present' essay required no such change. This does not suggest that there were no differences in the mental process required for the two essays, but that any such differences were not revealed by crude comparisons of group performance, or were obscured by other variables.

There seemed to be no differences in performance between essay prompts in the students' consideration of audience, unless we consider that being a pig rather than a bird or a fish is unpleasant because of the way people will relate to you. It is difficult to work out how much the perception of oneself as a pig is unpleasant as a personal image of oneself, and how much this perception depends on the attitude of others to pigs.

6.2.3.3 PW titles

Title aim - to elicit a persuasive essay that expresses ideas and gives reasons in order to persuade the reader to agree.

The results are shown in Table 25 (pretests) and Table 26 (posttests) below.

Table 25: Performance similarity on PW pretest titles

	F	p			
PW pretest titles	0.40	0.675			
			n	mean /15	SD
Violent films should/should not be shown on TV			22	7.318	1.836
People should/should not be forced to pay a fine for throwing rubbish on the streets			23	7.391	2.388
Alcohol should/should not be banned in PNG			23	6.913	1.564

There were no significant differences between performance on the pretest titles.

Table 26: Performance similarity on PW posttest titles

	F	p			
PW posttest titles	1.87	0.163			
			number	mean /15	SD
Right to choose (marriage partner)			22	10.000	2.024
Settlement in urban areas			23	9.609	1.644
Penalties for breaking road safety laws			23	8.957	1.821

There were no significant differences between performance on the posttest titles. The ordering, however, shows that the 'Right to Choose' essay produced the best essays and the 'Road Safety Laws' produced the worst. The title that seems likely to have aroused the greatest emotion and thus provoked the greatest personal involvement, judging by the performance results, was the one that the students tackled best. The role of emotion in the writing process seems, therefore, to be different when the writing type changes from narrative to persuasive. In contrast to narrative writing where the degree of pleasure associated with the experience was the telling factor, the best production of persuasive writing seemed to depend on how much the writer wanted to convince her readers of the opinion she wanted them to hold. This makes sense because in persuasive writing the writer is dealing with wishes for the future rather than memories of the past. The degree of emotional involvement will fuel the drive for the preferred outcome, which will, presumably, be associated with the prospect of a pleasant emotion in the future. The drive to achieve this will, therefore, help to

guard against the prospect of feeling bad at some later date. The results provide strong support for the view that having something to say is what drives the writing process.

All the titles seem to have been successful in eliciting an attempt at persuasive writing, although some essays contained narrative insertions, as noted above. The main cognitive requirements of the essay prompts seemed to be similar. In each case the student was presented with a statement and asked to agree or disagree, giving reasons. It seems that the titles required the students to imagine alternative outcomes depending on the position adopted, to compare them and present a preferred version. The effect of such requirements on individual students, however, seemed to depend on the level of writing development the student had reached. The inclusion of some narrative passages in persuasive essays was difficult to quantify because individuals differed within the group, because rates of progress appear to be uneven, and because there were too few instances of persuasive writing per student to make clear comparisons.

In the pretest persuasive writing prompts, no specific audience was stated. In the posttest prompts, students were asked to write their essays for *Post Courier* newspaper readers. It was felt at the time of the research design that the specific audience specification in the persuasive writing posttest prompts did not constitute a difference in audience consideration for the students, since the kind of persuasive argument they were asked to produce for both pretests and posttests was, in any case, the kind of writing that appeared in letters written to the newspaper on topics of current interest. The letters in the *Post Courier* would probably have been the students' only written models for persuasive writing. In addition to this, the students were told that all their pretest and posttest essays would be read and marked by teachers outside the school, so the real audience for the writing had been stated. It is clear, however, that there was a difference in audience specification between pretest and posttest. The audience specification for the pretests was implicit, while the audience specification for the posttests was explicit. This flaw in the research design may have made a difference to the students' performance.

The possible difference between pretest and posttest performance in persuasive writing as a result of the difference in audience specification has to be taken into account when interpreting the results. It may make a difference to the calculation of the amount of improvement the whole sample achieved in persuasive writing between pretest and posttest. For example, if the explicit audience specification made the posttest tasks easier than the pretest tasks, then the visible improvement in persuasive writing would be inflated. However, this possible difference will not be able to account for any differences in improvement in persuasive writing between the control group and the experimental group since both groups were subject to the same constraints.

One area of noted difference between individual students, which seemed to have occurred more in the pretests than in the posttests, was in the variety of ways students addressed their audience. Their relationship with their audience appeared to include one or more of the following:

- specification of themselves as 'I'
- removal of themselves to give impersonal argument
- direct address of audience using 'you'
- inclusive address of audience using 'we'
- implied inclusion of audience by frequent references to 'people' doing this that or the other, with the assumption that these observations would be shared

There appeared to be a greater switching between forms of address in some of the pretest essays than in the posttests. This could have been caused by the difference in audience specification, or by a lack of experience in persuasive writing, or by a mixture of both. It seems that essay prompts had not only differing effects because of the personality and preferences of different writers and possibly because of the difference between the implicit audience specification of the pretest compared to the explicit specification in the posttest, but also differing effects according to the level of writing proficiency of the subjects.

6.2.4 Summary

There were no significant differences in performance between titles, but differences in the order of performance on narrative prompts supported student comment that pleasant topics were easier to write about than unpleasant topics. Any differences in performance on the persuasive writing tests seemed to have been caused by the strength of emotional response that the essay topic evoked. The strength of the emotional response determined the amount of effort given to the persuasion. It seems that although there were individual differences in emotional response to specific topics which could not be predicted by the task, it should certainly have been possible to predict in a general way which topics were likely to evoke pleasant experiences and which were not. With persuasive essay prompts, it should have been possible to predict which topics were generally likely to arouse a passionate interest, and which were not. The results show that such considerations made a difference to performance.

6.3 Inter-rater reliability

6.3.1 Raters

The pretest and posttest scripts of all the writing types were marked by three raters, one male PNG non-native speaker and two female expatriate native speakers. The raters were experienced teachers in PNG. None of the raters knew the students nor to which group any student belonged.

6.3.2 Rating procedure

The raters were given the pretest scripts shortly after they were completed and received the posttests several months later. Each rater marked the scripts in his or her own time and the marking took several weeks to complete. This kind of marking reflected a natural marking situation where teachers break off from marking a set of scripts to attend to other matters and then return to the task when they can. Each script was marked out of a possible total of 5 (the highest mark), and then the three raters' marks were added together to give a mark out of possible total of 15. No attempt was made to bring the raters into line with each other, so that each script received the full value of three 'free

standing' evaluations. This method was employed deliberately to guard against possible problems of compromised validity (see Chapter 3.5.1.3).

The raters were provided with a holistic impression rating scale (see Appendix D) in order to standardise the levels of writing quality for the essays. The raters were all experienced in marking in such a way for this level of student writing and I discussed the scale with each rater before marking began. The scoring guide was described in 3.5.1.2 and required raters to judge each essay holistically with regard to certain criteria. It is possible that the imposition of these criteria made it harder for raters to rate holistically, but since no rater feedback on the rating scale was obtained, it is not possible to know exactly what effect it had. This was a shortcoming of the research design and would have been useful information to have.

6.3.3 Reliability figures

Pearson correlations were done to establish inter-rater reliability between the three raters.

6.3.3.1 PHN

Inter-rater reliability results for PHN are given in Table 27 below.

Table 27: Inter-rater reliability - PHN

Pearson Correlation	PRETESTS		
	RATER1	RATER2	RATER3
RATER2	0.599*		
RATER3	0.539*	0.524*	
pretest average	0.868*	0.840*	0.805*
* p = <0.01			
	POSTTESTS		
	RATER1	RATER2	RATER3
RATER2	0.358*		
RATER3	0.314	0.251	
posttest average	0.732*	0.755*	0.712*
* p = <0.01			
Key: Rater 1 - PNG non-native speaker			
Rater 2 - expatriate native speaker			
Rater 3 - expatriate native speaker			

(Everything above 0.3248 is significant at 0.01 for a two tailed test with 60 or more subjects.)

For both pretests and posttests there was a significant correlation between each of the raters and the average mark, but this should not obscure the fact that for the pretests correlations between pairs of raters ranged from .52 to .60, showing that they agreed on only 27 - 36% of the scripts (the variance overlap). For the posttests, the range was .25 to .36. Only raters 1 and 2 correlated significantly but this showed agreement on only 13% of the scripts. Raters 1 and 3, and 2 and 3 achieved consensus on very few scripts (6-9%).

6.3.3.2 ISN

Inter-rater reliability results for ISN are given in Table 28 below.

Table 28: Inter-rater reliability - ISN

Pearson Correlation	PRETESTS		
	RATER1	RATER2	RATER3
RATER2	0.486*		
RATER3	0.553*	0.498*	
pretest average	0.864*	0.769*	0.820*
* p = <0.01			
	POSTTESTS		
	RATER1	RATER2	RATER3
RATER2	0.377*		
RATER3	0.333*	0.293	
posttest average	0.715*	0.761*	0.755*
* p = <0.01			

Key: Rater 1 - PNG non-native speaker
Rater 2 - expatriate native speaker
Rater 3 - expatriate native speaker

(Everything above 0.3248 is significant at 0.01 for a two tailed test with 60 or more subjects.)

Once again, despite significant correlations of the raters with the average marks on both pre and posttests, the level of agreement between pairs of raters was poor. On the pretests, levels of agreement ranged from .49 to .55 (24 -30% shared variance), while on the posttests the range dropped to .29 to .38, with the greatest amount of shared variance only 14% (between raters 1 and 2), while raters 2 and 3 hardly managed to agree at all.

6.3.3.3 PW

Inter-rater reliability results for PW are given in Table 29 below.

Table 29: Inter-rater reliability - PW

Pearson Correlation	PRETESTS		
	RATER1	RATER2	RATER3
RATER2	0.558*		
RATER3	0.648*	0.549*	
pretest average	0.843*	0.846*	0.849*
* p = <0.01			
	POSTTESTS		
	RATER1	RATER2	RATER3
RATER2	0.385*		
RATER3	0.550*	0.475*	
posttest average	0.780*	0.789*	0.841*
* p = <0.01			
Key: Rater 1 - PNG non-native speaker			
Rater 2 - expatriate native speaker			
Rater 3 - expatriate native speaker			

(Everything above 0.3248 is significant at 0.01 for a two tailed test with 60 or more subjects.)
The correlations are significant at 0.01.

Once again a similar pattern emerged, where correlations between raters and average marks were significant, but levels of agreement between pairs of raters were low. On the pretests, the range was .55 to .65, but as with PHN and ISN, raters found it harder to agree on the posttest scripts (range from .39 to .55).

6.3.3.4 Summary of results

Despite the relatively low levels of agreement, there were significant correlations between most raters. Table 30, below, summarises the number of significant correlations between raters for all writing tests.

Table 30: Summary of significant correlations between raters

	PHN		ISN		PW	
	pre	post	pre	post	pre	post
rater 1 & rater 2	Y	Y	Y	Y	Y	Y
rater 1 & rater 3	Y	Y	Y	Y	Y	Y
rater 2 & rater 3	Y	N	Y	N	Y	Y

Y = Yes, significant correlation N = No significant correlation

Key: Rater 1 - PNG non-native speaker
Rater 2 - expatriate native speaker
Rater 3 - expatriate native speaker

(This table is based on the details of the Pearson correlations given in Tables 27-29 above.)

There were significant correlations between raters on all the pretests and on the persuasive writing posttests, but raters 2 and 3, who were both expatriate native speakers of English, did not agree on how to evaluate the narrative posttests of either type. It is not possible to know what caused them to evaluate differently from each other. It is worth pointing out that only months previously, when they evaluated the pretests, they had evaluated these types of writing in seemingly similar ways. At least they achieved significant correlations in PHN and ISN on the pretests, although the fact that there were significant correlations between their ratings does not necessarily mean that the raters evaluated in the same way. It means only that they reached similar overall marks on what is reckoned to be a significantly significant proportion of the scripts. Connor and Linton (1995) make the point that superficial agreement between raters may mask important differences.

The fact that the different evaluations of these raters were included and used in the overall rating and ranking of essays lends strength to the results, in that validity is increased. Lumley and McNamara (1995) make the point that the way raters evaluate can change over time, even when they have received rater training. Numerous researchers draw attention to the fact that validity can be compromised when pressure is put on raters to conform to a particular set of criteria or point of view (Charney 1984; Huot 1990; Vaughan 1991) so although raters were asked to score according to a rating scale with preset criteria (see Appendix D), their interpretations were their own. In view of the widely reported difficulty in obtaining inter-rater reliability without compromising validity as noted above, it is not surprising that the inter-rater reliability on the essays was low. Since overall

evaluations tend to rank rather than disclose specific qualities or poverties of writing on which those ratings are based, I prefer to draw attention to, and to rely on, the reliability and integrity of the raters as experienced ESL teachers rather than on the possibly false reassurance of statistically significant reliability figures reported here. At least we can be sure that the evaluations were done in good faith and all were included.

There is always a possibility that another set of raters would evaluate differently, or that the same set of raters might evaluate differently on another occasion. Awareness of the imprecision and valid disagreement contained in ratings should be kept in mind. We should be aware of this for the sake of the students whose lives we affect, as well as for the sake of being aware that the 'truth' about how writing works is dependent upon its readers, who may be more variable than we imagine. Having acknowledged this, it is hoped that the triple marking of each script used for the experiment has given a slightly more reliable result than the common double marking that is often used for examination purposes. It is hopefully, a slightly more reliable version of the kind of evaluation the students usually receive.

CHAPTER 7 - RELATIONSHIP BETWEEN THE WRITING TYPES

The first aim of the study was to investigate the relationships between the three writing types: personal history narrative, imagined story narrative and persuasive writing. Data from the pretests were used (i) to investigate the hierarchy of difficulty between the types using holistic impression ratings to calculate the Gutman scale, (ii) to compare objective differences on each of the three types, investigating grammatical structure, fluency and accuracy, and (iii) to test objective measures to see which of these were indicators of quality in the three writing types.

Two points need to be made at the outset. The first is that the relationship investigated was the relationship of the writing types at the beginning of the experiment. This means that the relationship between the types found here may be different from the relationship between the types of writing at a later stage of development. The second point is to draw attention to the mixing of writing types. Data from the student questionnaires and from an investigation of some of the essays (see Chapters 5 & 6) show that the writing types did not remain totally discrete. There were frequent inclusions of other types of writing within the main type.

Firstly, the results of the investigation into a hierarchy of difficulty will be reported. Secondly, there will be a comparison of the objective differences between the three types of writing. Thirdly, the 'good' and 'poor' scripts will be investigated to see which objective features discriminate between them. This will be followed by a discussion of problem essays which did not fit the usual profile associated with good scripts, and finally, there will be a brief summary of the findings.

7.1 Hierarchy of difficulty

It was expected that persuasive writing would be more difficult than ISN, which would in turn be more difficult than PHN. Gutman scaling was used to test for a hierarchy of difficulty using the impression marks given for each essay type. Scripts from the pretests were scored by holistic impression marking and were divided into satisfactory (10-15/15) or unsatisfactory (0-9/15). The scores were entered on an implicational scale. Subjects from the control group (34) and the

experimental group (34) were treated together so there were 68 subjects, who produced scripts aimed at each writing type. The results of the implicational scale are shown in Table 31 below.

Table 31: Implicational scale to show hierarchy of difficulty

Gutman Scaling for 3 essay types: PHN, ISN and PW - pretest data
Binary Scale 66.6%

	PHN		ISN		PW		
	0	1	0	1	0	1	TOTALS
	0	8	1	7	0	8	8
			....ERR...				
	10	9	0	19	19	0	19
	....ERR...						
	0	9	9	0	9	0	9
	...ERR...						
	32	0	32	0	32	0	32
SUMS	42	26	42	26	60	8	68
ERROR	10	0	1	0	0	0	10

C_{rep} (Coefficient of reproducibility) = 0.951 Coefficient of scalability = 0.931

The Coefficient of reproducibility (C_{rep}) means that 95% of the time it will be possible to predict a student's performance from his or her position in the matrix. For example, if a student is rated 'satisfactory' in persuasive writing, there is a 95% chance that the student will produce satisfactory pieces of ISN and PHN. However, the C_{rep} is calculated from the number of errors in the grid. To make sure the data is truly scalable (and to minimise the effects of artificial cut-off points) the coefficient of scalability has to be calculated. The coefficient of scalability is 0.931. This confirms that a hierarchy of difficulty exists between the three writing types. (Method of calculation and interpretation carried out according to Hatch & Farhady 1982.)

There are three points concerning the hierarchy of difficulty. The first is that the data used for the scale were holistic ratings, so in the light of the discussion on inter-rater reliability in the preceding chapter the results should be viewed cautiously. The second is to note that a scrutiny of the scale shows that the difference between persuasive writing and the narrative types was more stable i.e. had fewer exceptions to the overall pattern, than the difference between the narrative types of writing. This accords with intuition and previous argument (see Chapter 3) that although it was expected that

ISN would be harder than PHN, that a much more marked increase in difficulty could be expected with persuasive writing. The third point is a reminder that writing types were mixed, so that what is being compared are writing types that have a main focus on PHN, ISN or PW.

It is not possible to prove hierarchy of difficulty from an investigation of the average marks of each group of scripts, but it is worth noting that the pretest means were in accord with a difference in difficulty shown by the scale. The pretest means were:

		total 15 marks
PHN	-	9.12
ISN	-	8.99
PW	-	7.21

In addition, students made unsolicited comments that ISN was more difficult than PHN in the questionnaires they filled in after the project had finished.

The following hypotheses were confirmed by the results of the implicational scale:

1. Subjects who produce a satisfactory piece of persuasive writing will produce a satisfactory piece of imagined story narrative.
2. Subjects who produce a satisfactory piece of imagined story narrative will not necessarily produce a satisfactory piece of persuasive writing.
3. Subjects who produce a satisfactory piece of imagined story narrative will produce a satisfactory piece of personal history narrative.
4. Subjects who produce a satisfactory piece of personal history narrative will not necessarily produce a satisfactory piece of imagined story narrative.

7.2 Differences in grammatical structure, fluency and accuracy

This section will present data from the pretests to show the differences between the types of writing on the objective measures. Once again, it is necessary to remember that the sets are different in

overall writing function, but that each set can be expected to contain some mixing of writing types within it.

7.2.1 Grammatical structure

It was expected that there would be differences in grammatical structure between narrative types and persuasive writing to reflect the greater lexical density and complexity expected in persuasive writing, but it was not expected that there would be significant differences between the narrative types themselves. Contrary to expectations, the results, presented in Table 32 below, show some differences between all three types.

Table 32: Differences in grammatical structure

ANOVA					
	PHN	ISN	PW	F	p
n	(68)	(68)	(68)		
	mean	mean	mean		
t-units (per 100 words)	7.36	8.58	6.17	36.74	0.000*
words per t-unit	14.87	12.00	17.09	34.55	0.000*
error-free t-units (per 100 words)	3.15	4.40	1.54	57.12	0.000*
words (per error-free t-unit)	11.63	10.25	12.87	9.82	0.000*
* significant ($p<0.05$)					

The persuasive writing essays contained the longest t-units, as expected, while shorter t-units (just over 2 words shorter on average) were produced for the PHN category. Much shorter t-units again (another 3 words shorter on average) were written for the ISN category, and the error-free measures showed the same pattern. The fact that persuasive writing essays contained longer, more complex sentences came as no surprise since the structure of persuasive writing in relation to the structure of narrative writing has been well documented by Halliday and Hasan (1976), Halliday (1985) and others. The fact that there was almost as large a difference in t-unit length between ISN and PHN as there was between PHN and persuasive writing was unexpected. Two questions are raised by such a result. The first is to ask whether such large differences in t-unit length would occur between

narrative writing performances by the same subjects at other stages of writing development. The second is to ask why such a difference occurred.

The first question can be answered to some extent by comparing the number of words per t-unit of the pretests and the posttests for the narrative writing types. T-tests were carried out to compare means on the pretests and means on the posttests as shown in Table 33 below.

Table 33: Comparison of pretest and posttest t-unit length in PHN and ISN

	n	PHN mean	n	ISN mean	t	p
words per t-unit (pretests)	68	14.87	68	12.00	5.26	0.0000*
words per t-unit	68	12.142	68	13.078	1.67	0.098

*significant ($p<0.05$)

It can be seen that during the posttests the students wrote longer t-units for ISN than for PHN. There was a significant difference in length of t-unit between the narrative types in the pretests, but not in the posttests. The pattern had changed so it is clear that the relationship between the narrative writing types did not stay the same. As students developed competence in writing ISN, the structure of the narrative types drew closer together. The most likely explanation for the difference at the time of the pretests could be that when first faced with having to write invented stories, the students found it so difficult that they wrote very short stilted sentences.

7.2.2 Fluency

Differences in fluency between the writing types are shown in Table 34 below.

Table 34: Differences in Fluency

ANOVA							
			PHN	ISN	PW	F	p
n			(68)	(68)	(68)		
			mean	mean	mean		
no of words per essay (pretests)			275.84	268.60	232.72	4.62	0.011*
* significant (p<0.05)							
Scheff Test							
		CrDiff	X'i-X'j (Observed Diff)				
PHN	ISN	46.19	7.24		not significant		

It was expected that there would be differences in fluency to correspond with the differences in difficulty. The students were expected to be most fluent in PHN and least fluent in PW to correspond with expected levels of difficulty. The figures appear to support this expectation although the Scheff test[1] showed that the means of the narrative types were not significantly different. There is a significant difference only between PW and the narrative types., but not between the narrative types themselves.

7.2.3 Accuracy

7.2.3.1 Differences in number of errors

In overall accuracy there was a significant difference between persuasive writing and the narrative types, but not between the narrative types themselves. Most errors were made when writing persuasive essays, which was in accordance with expectations, and least were made when writing ISN, which was not in line with expectations. It had been expected that more errors would be made when writing ISN than when writing PHN, but this was not so. Table 35, below, gives the results.

Table 35: Differences in overall accuracy

pretest data ANOVA							
				PHN	ISN	PW	F	p<0.05
n				(68)	(68)	(68)		
				mean	mean	mean		
no of errors per 100 words				8.06	7.71	10.02	5.91	0.003*
Scheff Test								
		CrDiff	X'i-X'j (Observed Diff)					
PHN	ISN	2.49	0.35		not significant			

Although there was no significant difference in the frequency of error between the narrative types, more errors were made in PHN than in ISN and this was surprising. It is logical to suppose that there might be a direct correlation between the number of language errors and the degree of difficulty a piece of writing imposes. Since we have limited processing space, then a greater cognitive load could be expected to increase the frequency of error in our performance and there is evidence to support such a view (Bartholomae 1980; Peterson 1993). However, one strategy for dealing with difficulty, which mitigates against such a direct link, is for writers to err on the side of safety and not attempt unfamiliar and therefore risky forms of expression. Such a tactic would cut down the number of errors, although it might have the side effect of producing boring text. It is possible that the subjects found the ISN tasks so unfamiliar and difficult at the time of the pretests that they wrote short careful sentences in contrast to the way they wrote PHN. Such speculation accords with the finding that they wrote significantly shorter t-units for ISN than for PHN during the pretests. It is interesting to note that despite the Gutman Scale confirmation of a hierarchy of difficulty between the narrative types, which was supported by student comments, this was not reflected in differences between them on measures of fluency and accuracy.

7.2.3.2 Differences in types of error

The frequency of vocabulary errors and errors of reference differed between the types of writing. Errors with prepositions and omission and overall frequency of cohesion and coherence errors

[1] post-hoc comparisons of pairs of means

differed between persuasive writing and the narrative types, but not between the narrative types themselves. See Table 36 below for a summary of significant differences in types of error.

Table 36: Significant differences in types of error

pretest data/ errors per 100 words					
ANOVA					
	PHN	ISN	PW	F	p
n	(68)	(68)	(68)		
	mean	mean	mean		
Vocabulary	0.324	0.524	0.909	11.90	0.000*
Grammar					
prepositions	0.237	0.303	0.435	3.59	0.029*
Cohesion & Coherence					
reference	0.275	0.075	0.677	23.26	0.000*
omission	0.647	0.699	0.968	4.02	0.019*
Total	2.680	2.422	3.531	6.12	0.003*
* significant ($p<0.05$)					
Scheff Test					
		CrDiff	X'i-X'j (Observed Diff)		
Grammar - prepositions	PHN/ISN	0.20	0.07		not significant
C & C omission	PHN/ISN	0.34	0.05		not significant
C & C Total	PHN/ISN	1.16	0.26		not significant

(See Appendix I for a complete list of differences in types of error.)

There was a difference between writing types in frequency of vocabulary errors, defined as being errors where a wrong vocabulary item was used. This category was distinct from errors of spelling and from grammatical errors where the right word was used in the wrong form. The most frequent incidence occurred with persuasive writing, followed by a less frequent occurrence in ISN, followed by a further drop in PHN. Differences in levels of difficulty could account for this. Since production of PHN seems to be the easiest because the content is psychologically closest to the writer, then that would explain the comparatively low incidence of vocabulary errors in this type. Presumably PHN is the type of writing that most closely resembles oral language and where familiar vocabulary items would be used. ISN might be expected to need a larger and less familiar store of vocabulary than PHN, although this need not necessarily be the case. Persuasive writing lies at the most difficult end

of the scale because of its requirement to fulfil formal audience needs, which could be expected to require the kind of vocabulary that is not normally used for everyday purposes.

In the 'grammar' category there was a significant difference between persuasive writing and the narrative types in errors with prepositions, one of the most frequent types of error in PNG writing (Smithies & Hozknecht 1981; Phillip 1986). It is not clear why students made more of these errors when writing persuasive essays, fewer when writing ISN and fewest when writing PHN, except to suggest once again that the differing levels of difficulty made the difference. Maybe students used more prepositions for persuasive writing than they did for the narrative types or maybe the load on STM caused them to make mistakes they did not make when writing easier types of essays. There was, however, no significant difference between the writing types in the 'grammar' category overall.

In contrast, the average number of 'cohesion and coherence' errors varied between the types of writing where PW contained far more than the narrative types. In addition there were differences in the subcategories of 'reference' and 'omission'. Once again PW contained more such errors than the narrative types, although PHN and ISN differed on frequency of reference errors where PHN contained significantly more than ISN. This did not reflect expectations that ISN would be more difficult than PHN but was consistent with the overall lower level of error displayed by ISN essays in the pretests. The most noticeable difference at this stage of development was in the greater frequency of cohesion and coherence errors overall for PW compared to the narrative types. It seems that the coherence of text was particularly sensitive to the added cognitive difficulty imposed by persuasive writing.

7.3 Objective indicators of quality

'Good' and 'poor' pieces of writing in each type were investigated to find out which objective measures discriminated between 'good' and 'poor' scripts. The intention had been to use pretest scripts with ratings from 11-15/15 as 'good' essays, and those from 0 - 5/15 as 'poor'. Unfortunately these divisions did not yield a large enough sample of 'good' and 'poor' scripts for all the writing

types, so it was decided to use scripts with ratings of 10 or more as 'good' scripts, and scripts scoring up to 6 as 'poor' scripts. An unmatched t-test was then used to compare the means of the good and poor scripts on each measure to see if they differed significantly. The findings on these objective measures will be reported, followed by a brief discussion of essays which did not 'fit' the profile of objective measures normally associated with a 'good' or a 'poor' script.

When considering the results, readers are reminded once again that writing types are not expected to be discrete. There are two other points: one is the fact that the results rely in the first place on ratings of essays being reliable, and the second is the fact that samples of 'good' and 'poor' scripts were small. Inter-rater reliability was found to be dubious (see Chapter 6), but it is hoped that top and bottom scripts might be scored more reliably than those that bunched in the middle although such optimism needs to be tempered with caution.

7.3.1 PHN

It was found that both fluency and overall accuracy significantly influenced ratings of 'good' pieces of PHN produced by the grade nine students at the beginning of the study. A summary of those objective measures which discriminated between 'good' and 'poor' pieces of writing is given in Table 37 below.

Table 37: PHN - Summary of objective measures that discriminated between 'good' and 'poor' scripts (pretest data)

	good	poor	t	p
n	(26)	(9)		
Structure				
no of error-free t-units per 100 wds	3.86	1.86	3.61	0.0023
Fluency (average number of words)	330.4	174.6	6.73	0.0000
Accuracy				
category: Grammar	1.58	5.38	3.38	0.0097
Overall Accuracy(errors per 100 words)	5.35	13.72	-5.22	0.0005

The number of error-free t-units discriminated significantly between 'good' and 'poor' scripts. This was not surprising since we expect narrative writing to be characterised by grammatical intricacy rather than lexical density and the fact that it was the error-free measure of structure that discriminated supports Perkins' (1980) findings that it is the error-free measures that matter, at least at this level of proficiency in PHN. The fluency measure also discriminated between good and poor pieces of PHN. Writers of 'good' scripts wrote approximately twice as many words as writers of 'poor' scripts and the measure of overall accuracy discriminated, too. Writers of 'poor' essays made more than twice as many mistakes than the writers of 'good' narratives and errors in the 'grammar' category were the types which differed significantly. That grammatical error discriminated between 'good' and 'poor' scripts more than other types of error, emphasises the fact that readers' assessment of text seems to focus on form at basic levels of writing. At a stage where interesting texts are often not generated, the reader's attention appears to focus on the writer's command of standard English.

It seems at an early stage of writing development that what mattered most for the production of a 'good' piece of personal history narrative was length and a reasonable command of English. The narrative had to be long enough to make it interesting, and it had to be accurate enough, particularly with regard to word endings, for the reader not to be irritated by the mistakes.

7.3.2 ISN

Once again, both fluency and overall accuracy significantly influenced ratings of 'good' pieces of ISN. In contrast to PHN, there was no significant difference in the structure of the 'good' scripts compared with the 'poor' scripts, as measured by the number of error-free t-units. A summary of measures which discriminated between 'good' and 'poor' pieces of writing is given in Table 38 below.

Table 38: ISN - Summary of objective measures that discriminated between 'good' and 'poor' scripts (pretest data)

t-tests				
	good	poor	t	p
n	(26)	(8)		
Fluency (average number of words)	333.8	147	9.39	0.0000*
Accuracy				
*category - Cohesion & Coherence**	0.18	0.85	-4.09	0.0036*
Overall Accuracy (errors per 100 words)	6.19	12.68	-3.44	0.0074*
* significant ($p<0.05$)				
* see Appendix E for description of error category				

The fluency measure discriminated significantly between 'good' and 'poor' pieces in ISN. As with PHN, 'good' writers wrote significantly more than the 'poor' writers. The measure of overall accuracy, too, discriminated between 'good' and 'poor' pieces of writing. Writers of 'poor' ISN made nearly three times as many mistakes than the writers of 'good' stories, but the only specific error category which discriminated between them was 'cohesion and coherence'. It is possible that the level of difficulty associated with ISN caused weak writers to struggle to keep their text coherent. To be rated as a 'good' piece of ISN, what mattered most was fluency and overall accuracy where the writer was competent enough not to have too much trouble holding the text in mind and keeping it coherent.

7.3.3 PW

As with the narrative writing types, both fluency and overall accuracy significantly influenced ratings of 'good' pieces of persuasive writing. As far as structure was concerned the number of error-free t-units discriminated significantly between 'good' and 'poor' scripts. As well as overall accuracy, there were several particular types of error which marked a difference between 'good' and 'poor' scripts. Grammatical errors, vocabulary errors and errors of omission were significant discriminators. A summary of significant indicators is given in Table 39 below.

Table 39: PW - Summary of objective measures that discriminated between 'good' and 'poor' scripts (pretest data)

t-tests	good	poor	t	p
n	(8)	(25)		
Structure				
no of error-free t-units per 100 words	2.8	1.26	-3.42	0.0076*
Fluency (average number of words)	330.7	170.6	-5.64	0.0003*
Accuracy (number of errors per 100 words)				
Vocabulary	0.81	2.3	3.3	0.0025*
Grammar	1.61	3.03	3.64	0.0010*
Cohesion & Coherence	0.49	1.31	2.7	0.0130*
Overall Accuracy	7.18	12.31	3.99	0.0005*
* significant ($p<0.05$)				

As with PHN, the number of error-free t-units discriminated significantly between 'good' and 'poor' persuasive writing scripts, although I had expected that the number of words per t-unit would be the significant discriminator since persuasive writing is normally characterised by lexical density. It may be that at a later stage of writing development, the number of error-free t-units would no longer be a discriminating feature between 'good' and 'poor' scripts', but at this stage the number rather than the length of accurate stretches of text made a difference. Fluency, too, discriminated between good and poor pieces of writing. The writers of 'good' persuasive essays wrote approximately twice as many words as the writers of 'poor' essays, but made roughly half the number of errors. The measure of overall accuracy discriminated significantly and, as with PHN, the number of grammatical mistakes made a difference. Errors of vocabulary discriminated presumably because of the more formal and less familiar words that were required. In the 'cohesion and coherence' category, the errors that discriminated most were errors of omission. These were cases where students had missed out the main verb or some other item crucial for an understanding of the text. The load on STM was presumably so great that weak writers lost sections of text that they had meant to put in.

Once again, as with the narrative scripts, a 'good' piece of persuasive writing required a certain standard of fluency and overall accuracy. It mattered to be able to cope well enough with the load on

STM not to omit words that were needed for the meaning while at the same time maintaining a reasonable standard of grammatical accuracy.

7.3.4 Differences between writing types

At the subjects' stage of writing development at the time of the pretests, there seemed to be more similarities between the objective indicators of quality in each type of writing than differences. The amount of fluency and overall accuracy was associated with quality in all three types. The main difference was that a greater number of error categories discriminated between 'good' and 'poor' persuasive essays, than between 'good' and 'poor' essays of either of the narrative types.

7.3.4.1 Grammatical structure

The number of error-free t-units discriminated between 'good' and 'poor' pieces of writing in PHN and PW. This was perhaps surprising in the case of persuasive writing, where lexical density could be expected to be a significant feature. It seems that, although students wrote longer t-units for persuasive writing than they did for the narrative types, this measure was not a significant indicator of a 'good' piece of PW.

7.3.4.2 Fluency

The fluency measure discriminated significantly between 'good' and 'poor' essays in all three writing types. The results are shown in Table 40 and Figure 1 below.

Table 40: Fluency measure as an indicator of quality

unmatched t-tests average number of words per essay (pretest data)				
PHN	good	poor	t	p
n	26	9		
	mean	mean		
	330.4	174.6	6.73	0.0000*
ISN	good	poor	t	p
n	26	8		
	mean	mean		
	333.8	147	9.39	0.0036*
PW	good	poor	t	p
n	8	25		
	mean	mean		
	330.7	170.6	-5.64	0.0003*
*significant ($p<0.05$)				

Fluency was associated with quality in all writing types. It is interesting to note that the 'good' scripts in all three writing types were almost identical in length, achieving an average of 330 words. These 'good' scripts were at least twice as long as the average 'poor' scripts in each type. Lack of familiarity with a particular type of writing did not seem to have made a difference to the attempt to tackle it in some way or other, sometimes presumably by mixing in types of writing that could be managed more easily. It seems that the students dived onto the page any way they could.

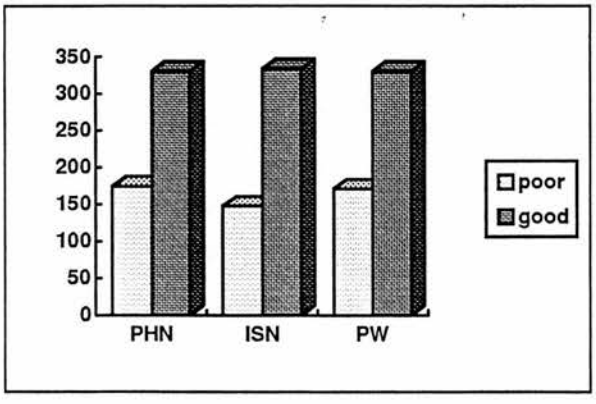

Figure 1: Fluency measure as an indicator of quality

7.3.4.3 Accuracy

The measure of overall accuracy discriminated between 'good' and 'poor' pieces on all three types of writing. 'Good' writers made significantly fewer errors than 'poor' writers. The results are shown in Table 41 below.

Table 41: Overall accuracy as an indicator of quality

unmatched t-tests average number of errors per 100 words (pretest data)				
PHN	good	poor	t	p
n	26	9		
	mean	mean		
	5.35	13.72	-5.22	0.0005*
ISN	good	poor	t	p
n	26	8		
	mean	mean		
	6.19	15.94	-3.44	0.0074*
PW	good	poor	t	p
n	8	25		
	mean	mean		
	7.18	12.31	3.99	0.0005*
*significant ($p<0.05$)				

The 'good' essays in each of the writing types showed a level of error that was roughly half that of the 'poor' essays. The 'good' persuasive essays contained more errors than 'good' ISNs, which in turn contained more errors than the 'good' PHNs. This probably reflects the fact that the students already had some experience in PHN, but not in the other two types of writing. If level of accuracy is an indicator of difficulty, then the performance of 'good' writers supports the hierarchy between the writing types. This finding contrasts with the comparison of amount of error overall between ISN and PHN where more errors were made in PHN (see 7.2.3.1). Since both 'good' and 'poor' writers of ISN made more errors on average than the 'good' and 'poor' PHN writers, it must have been the middle range of ISN scripts that contained the careful relatively error-free text. Table 42 below shows which error categories discriminated between 'good' and 'poor' scripts in the writing types.

Table 42: Error categories as indicators of quality

unmatched t-tests to compare 'good' versus 'poor' scripts in each type (pretest data)

	good	poor	Error Category	t	p
PHN	26	9	Grammar	3.38	0.0097*
ISN	26	8	Cohesion & Coherence	-4.09	0.0036*
PW	8	25	Vocabulary	3.3	0.0025*
			Grammar	-3.64	0.0010*
			Cohesion & Coherence	2.7	0.0130*

* significant ($p<0.05$)

It is interesting to note that a lack of errors in the 'cohesion and coherence' category were indicators in ISN and PW, but not in PHN. This finding supports the view of the increased cognitive difficulty of ISN and PW compared to PHN where poor writers presumably failed to make their writing coherent because of an increased load on STM. Contrary to the expectations of many PNG teachers, spelling errors did not discriminate significantly between 'good' and 'poor' pieces of writing in any of the types investigated.

I would like to emphasise that the measures investigated were not intended to provide a comprehensive explanation of the differences between pieces of writing considered 'good' and those considered 'poor'. Considerations such as the level of interest called forth in a particular reader or group of readers by the effects of imagery, rhyme and rhythm, for example, are acknowledged but are beyond the scope of this research. The next section will investigate some scripts which were rated as 'good' pieces of writing, but which did not conform to the profiles, reported above, that are usually associated with 'good' scripts.

7.3.5 Problem essays

The limitations of the results given above need to be emphasised. Quantifying objective measures made two assumptions:

a) that the degree of quantity or absence of a feature was its most significant attribute.

b) that the measures chosen were significant indicators of the quality of writing

Neither of these assumptions can be proved but the second assumption is illuminated by a scrutiny of some of the 'good' scripts, which showed exceptions to the broad generalisations which is all that an investigation of some common features can attempt. The fact that the features chosen for investigation are those which are common concerns of teachers should not blind us to other factors which have not been considered. The first problem with using the objective measures reported above is that they are limited. They do not access those qualities of interest, novelty, rhythm or image: the qualities that make writing memorable, the 'magical' qualities of writing. The second problem is the assumption that quantity matters. For example, the fact that the difference between the number of grammatical errors is a significant discriminator between 'good' and 'poor' PHN, whereas the difference in the number of spelling errors is not, means only that there is a bigger difference in the number of grammatical errors. It seems likely that more equals more important but this may not necessarily be so. Other differences between 'good' and 'poor' scripts, which might not be revealed by statistical analysis, could be important. The third problem is that the method of analysis obscures the problem essays, the ones that 'don't fit'. To illustrate this problem, there follows a brief discussion of a 'good' essay from each writing type which did not fit the profiles described above as being normally associated with a good piece of writing in that type.

7.3.5.1 A PHN example

Subject 26

A BRIDE PRICE CELEBRATION I WILL ALWAYS REMEMBER

(Took Place in Milne bay Province)

A bride price celebration in my village was really interesting. there were not many people there becasue the population in the village was very low. It was that the woman in the village got married to a man living in a island far away from our village. and this man sail all the way to the our village to pay for the bride, there were alot of things which he also brought, they were, Clay pots, necklaces made of shells, tapa cloths head dresses. The people on the woman's side exchange their goods with the bride's husband. After that the village people sang and danced, In the night the people from near by villages came to celebrate as well, there were lots and lots of people out. There were also people who were specialised to carry out a particular job.

The girls were told to fetch water from the river and the boys were told to carry food from the garden, get firewood and also kill the pigs. In the evening, the boys made the fire and the girls peel

the vegetable and some other boys cut the pig, about (9.00) nine o'clock in the night, all the people in the village and the visitors had their meal. after the meal the elders, got into their traditional costumes. and started celebrate. The dance continue till morning, the people fed the visitors with left over pork and vegetable, after the breakfast the elders told the small boys to push the sailing canoes to the sea and then the village people made a farewell song and then the visiotrs got onto their canoe and sail out to the sea

Some of the village people were sad and some were cried especialy the women because they knew that is was the best bride-price celebration.

And myself too. I really like it best and it was very wanderful because I've never seen a bride.price like this before the things they use were mostly in traditional methods and they didn't use money to pay the bride and also it is memoriable to me because I found many friend from different villages and also there was trouble but only singing and dancing.

Result - rated as a good essay
Impression mark 11/15 9 errors per 100 words (cf 5.35 - good group's average)

This essay did not fit the profile of a 'good' personal history narrative essay from the point of view of accuracy. It had far more errors than the 'good' scripts normally had, yet somehow the writer has managed to communicate his emotion and has made us believe in it ` *it was very wanderful...*'. He has made us share in his good feelings at the memory of the brideprice he describes. Hamp-Lyons (1991e) makes the point that when an essay fails to engage the reader, it is at that point that mistakes are noticed. The opposite is obviously true too. This essay engaged the reader from the first words, ' *A bride price in my village was really interesting..* ' The reader's anticipation is aroused by this, so he or she reads on to find out what was 'really interesting' about it.

Although the essay did not fit the profile from the accuracy point of view, it did display the fluency normally associated with 'good' PHNs. The fluency feature consisted merely of a word count, but its significance is probably explained by the fact that more words mean a longer story and more details. The fact that fluency mattered was confirmed by the fact that some of the 'poor' PHNs, which did not fit the profile, failed because they were too short. These short essays were frequently accurate in that they contained virtually error-free language, but they were not long enough. It seems that one of our expectations of a story is that it does not finish too quickly. We like to 'get into it', to live in it and to

do this we need a certain amount of length and some detail. Then it becomes pleasurable. For example, if the above account had been something like 'A man from a far off place married a woman from our village. I enjoyed it because we sang, danced and ate a lot and I made some new friends', it would have been accurately written from a language point of view, it would have been a fairly accurate summary of the main content, but would have missed two vital features: the excitement of the writer's account, the emotion, and the images provided by the details.

The excitement in the account seems to be communicated by the rhythm of the sentences. For example '*In the evening, the boys made the fire and the girls peel the vegetable and some other boys cut the pig, about (9.00) nine o'clock in the night, all the people in the village and the visitors had their meal. after the meal the elders, got into their traditional costumes. and started celebrate. The dance continue till morning.....*' and so it goes on. There is a breathlessness in the writing. The writer cannot get it all out fast enough, so despite the fact that coherence sometimes suffers a little through the rapid fire of one piece of information after another and the punctuation is often dubious, the rhythm carries the reader on in a rush. It is like a piece of music. Consider for instance: '*....., there were alot of things which he also brought, they were, Clay pots, necklaces made of shells, tapa cloths head dresses.....*' The capital letter given to 'Clay' signals a new sentence or at least a proper noun, but this signal is overridden by the power of the narrative rhythm taking us forward.

It is true that none of the objects or actions are described in full detail. We are not told of colours or shapes, but some of the images still hang in the mind like pictures.

Consider:

'...after the breakfast the elders told the small boys to push the sailing canoes to the sea and then the village people made a farewell song and then the visiotrs got onto their canoe and sail out to the sea

You can almost hear the song and see the canoes sailing out to sea. The power of images, especially emotional images like the farewell song followed by the boats sailing away, imprint on the memory and involve the audience. The storyteller's emotion is important. If the emotion is left out of the

narrative, much of the interest is lost. It is what helps make a story interesting and interest is what carries us forward in the text and makes us 'forgive' a certain amount of inadequacy in other areas.

Rhythm, too, gives a sense of excitement and involves the reader. The events come pounding out thick and fast. The super text is 'God, it was good!' and the reader relates to this, feels good, wants more and reads on. This emotional message comes through the rhythm of the text. De Beaugrande (1982) points out that interest is essential in the judgement of a good story and is vital for motivation and memorability.

Another feature which is recognised to be important to the quality of a text (e.g. by researchers such as Halliday & Hasan 1976, Bamberg 1983) which was not investigated by the study, was text organisation. This piece of writing was organised in the following way:

beginning - announcement of an interesting brideprice & introductory background

middle - description of the brideprice celebration

end - reflection by people who had taken part, authors' reflection

This organisation satisfies the criteria noted by Cortazzi (1994) to be crucial for a successful narrative: the beginning, a state of equilibrium; the middle - a state of tension and change; and the end - a resolution or outcome. It may be the case that the good text organisation helped to outweigh considerations of accuracy.

It is possible that the novelty factor supplied by the fact that the bridegroom came from an island 'far away from our village' contributed to the text quality. Barritt, Stock and Clark (1986) found that raters valued novelty or surprise in the writing. There were plenty of details, too, and it is the details in stories that give much of the pleasure, because they somehow make the experience feel real, easy to relate to. When reading '*....the people fed the visitors with left over pork and vegetable....*' you become aware of the large amount of food which had been prepared for the gathering so that there was still some left over for breakfast. You remember how nice it is to eat again after staying up all

night, and you remember occasions when you have done the same thing. You contemplate the pleasure of eating left- over pork and vegetables in the open air as the sun comes up.

The fluency measure is obviously important. It seems clear that in this essay fluency overrode considerations of accuracy. Most of the 'poor' PHNs were accurate, but too short, as mentioned above. Unfortunately, the fluency measure in this study was as crude as it could be - number of words per essay. It would be interesting to investigate the effect of sound and rhythm on readers. The power of rhythm in this case seems to have communicated the emotion and through it, the atmosphere that seems to be crucial to all good stories. The number of words seems a pathetic measure to capture such crucial features, but it did seem to work in a rough way, and it did seem to be a more important measure than accuracy.

7.3.5.2 An ISN example

Subject 4

A DAY IN THE LIFE OF A BIRD

Iam a bird with long feathers with very beautiful, different colours like the rainbows. I feel proud of myself too. I live in a far away jungle where nobody can hurt me. I fly from brench to brench visiting my dearest friends and saying Goodmorning or Hellow to each other, I have two brothers and two sisters. as well as my father and Mother We all live together in the top breanch of a high tree were nobody can hurm us. I go out and play games with my brothers and sisters near the bush were there are planty of things to play with. On One Sunny morning I dessided to go out and visit the outside world on my own. I flew up above the trees and up into the sky and head for what I wanted to do. While I was flying I saw too many people on the street walking heare and there, Some going to the beach and Some going across the road by truck, others were just mainnig their own business. As I was flying, I decorved So many things that I didn't knew about. My arms felth tired and I decided to rest for a whie, I flew over and set on a brench of a tree were no body could see me. I felth the cool brezze and forgot about the all thing and went to sleep. while sleeping, rain fell and wet me, I got up and opean my eyes to see the jungle, but to my supprice I sew the great big city, just right in frount of me. I realised that I had come to visit the opean world and felth tired and went off to sleep. So I flew up above the city into the sky and head for my home in the far away jungle, I flew with great happiness inside of me, and Marking different styles in the air as I was flying. At last I made it home. My parrents, brothers, and sisters asked me, where I have gone. I got up and told them the all story. My father got up and said, You are a very brave bird and very clever. I shell never forget your

bravenest, till I die, because I've never gone out to the outside world to see it. And so I'll call you, "the brave one" I felth Proud of Myself and thank my father for that.

Result - rated as a good essay
Impression mark 12/15 9.8 errors per 100 words (cf 6.19 - good group's average)

The function of ISN is to invent experience in order to play and to learn and the content of this story fulfilled this function superbly. Like the personal history narrative discussed above, it satisfied the fluency criterion which is normally associated with a 'good' ISN, but failed from the accuracy point of view. So why was it rated as a 'good' essay? What was it that made the rater forget the mistakes and enjoy reading it?

Text organisation did not form part of the study and yet this essay shows that the text organisation was clear and satisfying. The bird was introduced, flew off on a dangerous adventure and returned safely. There was a beginning, a middle with dynamic change and tension and a resolution. These are Cortazzi's (1994) criteria for narrative satisfaction, which were mentioned in connection with the personal history narrative discussed above.

The degree of audience involvement was not assessed and yet it is clear that this is an obvious factor in whether or not an essay achieves a good rating. Audience involvement was one of the most obvious strengths of this essay. It shared the writer's emotion of pride and pleasure at the achievement of having left home for a dangerous adventure and returned safely. It created feelings of pleasure with sentences like: *I flew with great happiness inside of me...* Wouldn't we all like to fly with great happiness inside of us? It communicates feelings of escape and pleasure. The story content is enjoyable from the very beginning because it starts off with an expression of the writer's pleasure of being. What more could one want than to be a bird with long rainbow coloured feathers? The writer considers herself to be beautiful. Feeling beautiful makes everyone feel good, makes relationships likely to be good, opens up possibilities of living that feeling bad and ugly prevents. Grace Nichol's (1993) 'The Fat Black Woman's Poems' have the same effect. They show the

pleasure of liking the body you live inside, which makes life a delight, and so we want to read on. None of the measures used in the study allows for the importance of content and the need for the reader to identify with the hero of the story. Content and emotional identification of the reader with the characters in the story is clearly important.

The power of the images helps the essay to be pleasurable and memorable. Two images in particular stand out. One is the picture of the bird with rainbow coloured feathers. The second is the image of flying loop the loops and other patterns while flying home: '*I flew..... and Marking different styles in the air as I was flying*'. Once again, none of the objective measures accessed the imagery that can make texts memorable.

Provision of detail is enabled by the fact that the story is not too short, and provision of detail is important. It makes the story real to read that the bird fell asleep in a tree and '*while sleeping, rain fell and wet me..*'. Making the story feel 'real' allows the reader the vicarious experience. This is illustrated by the fact that a summary of the story content would not be interesting or pleasurable: 'I was happy at home in a safe place, but then risked a dangerous flight to look at the big city, and finally got back home safely.' The details have gone.

Sound and rhythm play a part in carrying the reader forward. Consider: '*I fly from brench to brench visiting my dearest friends and saying Goodmorning or Hellow to each other,...*' If you read this aloud, you can hear the pauses after 'Goodmorning' and 'Hellow', which makes us feel that we are actually saying or hearing the familiar greetings. We may not be aware that we hear the words that we read, but research (Snowling 1985) shows that it is not possible to process the written word without an auditory association. Even deaf people have some auditory image in the head for every word. It does not matter if our auditory image is 'wrong', e.g. we might pronounce 'picturesque' as 'pictureskew' in our heads, but it matters that we have some sound in our head for each word. Without the sound image we cannot process or ever remember a word. It is part, too, of our common experience that series of sounds, or rhythms have certain emotional messages that we all decode in a

similar way. For example, we all recognise when music has a peaceful rhythm or when it is exciting. If we hear sounds in our heads to go with the words we read, then it follows that we also experience series of sounds. The rhythms of texts, even prose texts, presumably give us messages that we may not be aware of, but which can be expected to be powerful nevertheless.

The consideration of the power of rhythm, image, content and detail seem to be important features that should be taken into account when considering what makes readers assess stories as 'good', and none of these features were investigated in the study. The measure of fluency merely provides the space to allow these features to exist, which goes a little way to explaining why fluency was a more powerful measure than accuracy, but also why it was too unspecific to mean much by itself. Accuracy was important, it seems, until overshadowed by content, rhythm and imagery, while a certain amount of fluency, or length, was always necessary because short texts are not experienced as stories.

7.3.5.3 A PW example

ALCOHOL SHOULD BE BANNED IN PNG

Alchol should be banned in PNG because of many certain reasons and also it is dangerous to health..

Firstly I'd say it is bad to health because drinking the alchol often is quiet serious, it will burn the heart badly because of the dangerous drugs used in it - and it also occupies alot of space in the stomach so the food doesn't settles properly in the stomach - That is how people go vomiting out all the stuff that they drink with the alchol

Secondly drinking alchol will also cause families to tear apart as well and also cause families to face financial problems Eg: The man uses up all the money to buy alchol and drinks with his friends and enjoyes himself, after drinking he comes home and beats up the wife and chilren, The wife then packs up and leaves him while he is heavily drunk - when the weekend is over, he goes to the kitchen and finds breakfast - Nothing is found, he checks his bag not even a coin is inside - He doesn't go to work because of financial problem he faces, at the end of the week, He recieves a letter from he's boss saying - DON'T BOTHER COMING BACK.

Thirdly alchol should be banned in PNG becasue at these stage - The New Generation starting from the young teenagers to early adults, everyone of them are taking alchol - They drink alchol in groups and are tempted to rape, break and enter and steal as criminals - when they are caught they end up in the jail belted by the police: I also believe that people at this stage shouldn't drink alchol because they don't think fo their future - some of them are JOBLESS just hanging around and

drinking from other's money. Even some of them just involve in little fight and starts it up as a MAJOR fight - I am standing on my point - Alchol is especially dangerous to teenagers. their Lungs will be burned up quicky as a piece of paper burning up by the hot gass of the alchol and they will end up at NINE MILE, six feet low:

People drinking alchol shouldn't be allowed to drive too! because they are made half sensed by the alchol.

Lastly alchol is RUBBISH and it will only SPOIL OUR LIVES and
RUIN YOUR GOOD CHARACTER..

- YOUR GOOD BUILT

- YOUR GOOD FRESH MIND.

Result - rated as a good essay

Impression mark 10/15 10.9 errors per 100 words (cf 7.18 good group's average)

The function of persuasive writing is to persuade so it follows that a 'good' essay needs to contain not only a cogent argument, but an effective presentation of that argument. The ideas and opinions have to be presented in a way that readers can relate to, follow, and be convinced. Somehow the text has to arouse and keep our interest. The persuasive essay above flouted conventional expectations of the genre by introducing speech, and pseudo shouting through the use of capital letters, but the raters obviously forgave these oddities because the text aroused interest. It feels as though the writer is shouting his message, that he is addressing you directly and insistently so he is difficult to ignore. In this way he forges a relationship with the reader that makes it difficult to get away or to ignore the message. It is an excellent essay from the point of view of audience involvement.

Once again, the fluency feature associated with 'good' writing was present, but the accuracy was lacking. What is interesting in the case of all three essays that did not fit the norm of linguistically accurate text, is to ask at what point the liveliness of the writing, generated by its rhythm, and the power of the emotion, enabled by the images and the content, override the need readers perceive for text to be accurately written? In the case of persuasive writing, it is interesting to ask how much the force of emotion affects perception of the argument and makes readers forget inadequacy or inaccuracy in the logic.

7.4 Summary

I would like to emphasise that the findings on the relationships between the writing types were those found at an early stage of writing development. Readers are reminded that the subjects had some experience in personal narrative, but little or none in persuasive writing and imagined story narrative at the time of the data production. A hierarchy of difficulty, as hypothesised, was found to exist where PW was more difficult than ISN, which was in turn more difficult than PHN. The results of the Gutman implicational scale were supported by student perceptions reported in Chapter 5.

Relationships between the types of writing, described according to the performance of the whole cohort, showed some differences. Structure, as measured by number and length of t-units, differed between the types. Persuasive writing had the fewest and longest t-units, while ISN had the most numerous and the shortest ones. The structure differed significantly between PHN and ISN but an investigation of posttest data showed that, after the writing practice, the differences in t-unit structure between the narrative types had almost disappeared. It seems that the relationship between writing types depended partly on the level of writing maturity in particular types of writing, rather than on some holistic competence in writing possessed by a writer at a particular point in time. The number of vocabulary errors differed significantly between each type of writing, where PHN had least errors, ISN had more and persuasive writing had most. Fluency was different between the persuasive and the narrative types, where persuasive essays were shorter than narrative essays, but there was no significant difference between the two narrative types. Overall accuracy, too, was shown to be different between persuasive and narrative, but not between the narrative types. Persuasive essays contained more mistakes than the narrative essays.

Indicators of good essays were fluency and overall accuracy for all three types of writing. A lack of grammatical error was an indicator of quality for PHN and PW. A lack of error in the 'cohesion and coherence' category marked 'good' pieces of both ISN and PW. It seems that writers struggled with the load on STM in these types of writing so that they tended to make careless errors and errors where items important to the sense of the text, e.g. main verbs, were omitted. Good PW essays were

further characterised by a lack of vocabulary errors. These were instances where students had used wrong or incomprehensible lexical items. In contrast, a lack of spelling errors was not an indicator of quality for any of the types. The most powerful indicators of quality in all three types were fluency and overall accuracy, although an examination of essays that did not 'fit' the normal profile associated with 'good' essays found that fluency seemed to be a more powerful indicator than accuracy. It seemed to be more important that an essay was long enough than that it was accurate. Other factors such as interest and audience involvement seemed able to override problems of accuracy, but an essay could not be rated as good if it was too short. Fluency seemed to be important because it provided the detail necessary for involving the audience and audience involvement was a clear strength in those essays which were scrutinised closely. Other textual features played a part too: rhythm, sound, imagery. Such features were outside the scope of this study, but their power has to be acknowledged.

CHAPTER 8 - THE DEVELOPMENT OF WRITING COMPETENCE

The second aim of the research was to investigate how students developed competence in each of the writing types over three quarters of an academic year i.e. three out of four terms. When considering the results, readers are asked to keep in mind the issues discussed in Chapter 6, particularly that a mixing of writing types was a feature of development during the treatment and that inter-rater reliability with an overall range of .49 to .65 on the pretests and .25 to .55 on the posttests was not high. This chapter will report on the writing development that took place between the pretests and the posttests.

8.1 Holistic ratings

Results show that the students' writing improved in all three writing types. T-tests between pre- and posttest results were carried out to show differences and these are presented in Table 43 below.

Table 43: Improvement over time on PHN, ISN and PW

	n	pretest mean(/15)	posttest mean(/15)	change	t	p
PHN	68	9.09	11.03	+1.82	-5.77	0.0000*
ISN	68	8.96	10.72	+1.77	-5.47	0.0000*
PW	68	7.21	9.53	+2.32	-7.01	0.0000*

*significant ($p<0.05$)

The results show a significant difference between pretest and posttest performance. The whole cohort became more skilful in all three types of writing despite the fact that the control group received no practice in ISN, the experimental group received no practice in PHN and neither group received practice in persuasive writing. It seems probable that the mixing of writing types that took place during the treatment can partly account for the improvement. (See Appendix J for samples of pretest and posttest essays.)

The fact that there was a large improvement in persuasive writing suggests that the explicit audience specification on the persuasive posttest prompts may have made the tasks slightly easier although it

is not possible to be sure. It seems unlikely, however, that this would account for such a significant increase in the average marks. The other explanation is that some of the mental operations necessary for persuasive writing, such as the evaluation of ideas and their likely consequences, were being practised in the narrative types. Consider the following extract from a treatment essay (The first time I watched television):

...I thought that television was a harmful thing but some people like me think that it is harmful because it hurts people. Even though it is not harmful but it is harmful because it causes rascalism[1] just by gaining ideas from it.Last but not the least, afterwards, I though that it was a good source of entertainment for showing informative and educational ideas towards the students but on the other hand, it was a bad source when it showed dirty and rascalism entertainment but as fas as Iam concerned television is good...'

This shows that the student has gone beyond his personal experience, which was a painful memory of humiliation because he had tried to catch a cricket ball batted by a player on the TV and his uncle had laughed at him, to consider the effects of television generally. He considers good aspects and bad ones and comes to a decision, such as he would have to do in a persuasive essay. Since neither group received practice in persuasive writing, and since each group received teaching in only one kind of narrative, the results seem to indicate that practice in both narrative types helped performance in narrative writing generally and also aided the transition to persuasive writing.

8.2 Change in objective measures between pretests and posttests

Since there was evidence that development of the students' writing competence had taken place in all three writing types, it was interesting to see what objective changes had taken place. T-tests were carried out to determine changes between the pretests and the posttests in grammatical structure, fluency, degree of accuracy and type of error.

8.2.1 PHN

8.2.1.1 Grammatical structure

The number and length of t-units were compared to see whether they were becoming longer or shorter or remaining stable as writing competence increased. The results are given in Table 44 below.

Table 44: PHN- Change over time in grammatical structure

t-tests n=68	pretest mean	posttest mean	change	t	p
Grammatical structure					
t-units (per 100 words)	7.36	8.51	+1.15	3.68	0.0002*
words per t-unit	14.87	12.14	-2.73	-4.90	0.000*
efts (per 100 words)	3.14	4.51	+1.37	4.94	0.0000*
words per eft	11.46	10.4	-1.06	-2.15	0.033*
*significant ($p<0.05$)					

Grammatical structure changed significantly as writing competence developed in PHN. The number of t-units increased and their average length became shorter. T-unit length seemed to settle down by the time of the posttest to round about 12 words on average. Error-free measures showed the same pattern. This could be interpreted as a settling down into a storytelling style that felt comfortable as opposed to a less controlled splurge of words at an earlier stage of development. It is probably a result, most of all, of an increased control of sentence structure. It is slightly surprising that the students started off writing longer t-units which decreased as they became more skilful. We might have expected students to write shorter t-units at an earlier stage of narrative writing, which then became longer, even if the clause lengths still remained shorter than those produced for persuasive writing. Long t-units are associated with a more complex, more lexically dense written style. What probably happened was that the t-unit length was longer at an earlier stage of writing development because the students were having more problems with sentence structure at that point.

[1] criminal behaviour

8.2.1.2 Fluency

The results of change over time in fluency for PHN are given in Table 45 below.

Table 45: PHN- Change over time in fluency

t-tests n=68					
	pretest mean	posttest mean	change	t	p
Fluency					
av. number of words per essay	275.8	377.5	+101.7	6.31	0.0000*
*significant ($p<0.05$)					

The change between the pre and the posttests in the fluency of the writers was significant. On average the subjects wrote a third more in the posttests than they had in the pretests, a hundred words more in an hour than they had written only a few months previously. The results of the investigation into features associated with quality in writing, reported in the previous chapter, showed that one of the most significant discriminators between 'good' and 'poor' scripts was the number of words the subjects wrote for their essays. During the pretests the writers of the 'good' essays had written on average 330 words per essay. By the time of the posttests the whole cohort was writing an average of nearly 380 words per essay. One of the most significant markers of the development in the writing competence of PHN was an increase in writing fluency.

8.2.1.3 Accuracy

Change over time in overall accuracy and in types of error for PHN are summarised in Table 46 below.

Table 46: PHN- Significant change over time in accuracy measures

t-tests measures= average no of errors per 100 words	pretest mean	posttest mean	change	t	p
Accuracy					
Vocabulary	0.32	0.54	+0.22	-2.66	0.0087*
Grammar:					
article	0.61	0.35	-0.26	2.73	0.0075*
redundancy	0.26	0.14	-0.12	2.59	0.011*
Total	3.78	2.83	-0.95	2.67	0.0086*
Cohesion & Coherence:					
reference	0.28	0.09	-0.19	3.47	0.0008*
punctuation	1.61	0.67	-0.94	4.87	0.0000*
Total	2.68	1.64	-1.04	3.99	0.0001*
TOTAL ERRORS	8.06	6.23	-1.83	3.04	0.0029*

*significant ($p<0.05$)

(Please see Appendix K for a complete list of changes over time on objective measures.)

PHN showed a significant decrease in the number of errors between the pretests and the posttests. The average number of mistakes reduced by almost a quarter. The number of errors decreased in some types of grammatical error, article errors and errors with redundancy, as well as in the category overall. In the 'cohesion and coherence' category, reference errors and punctuation errors dropped significantly. Vocabulary errors, however, increased as the students became more competent. It is worth noting that the only category of error that was a significant indicator of 'good' PHN scripts at the time of the pretests was a lack of grammatical mistakes and yet when the whole cohort was observed over time, there was a marked drop not only in grammatical errors, but also in the category of errors to do with cohesion and coherence. What is most interesting, however, is to see how the *proportions* of error changed over time. Figure 2 below shows how these changed as writing competence developed.

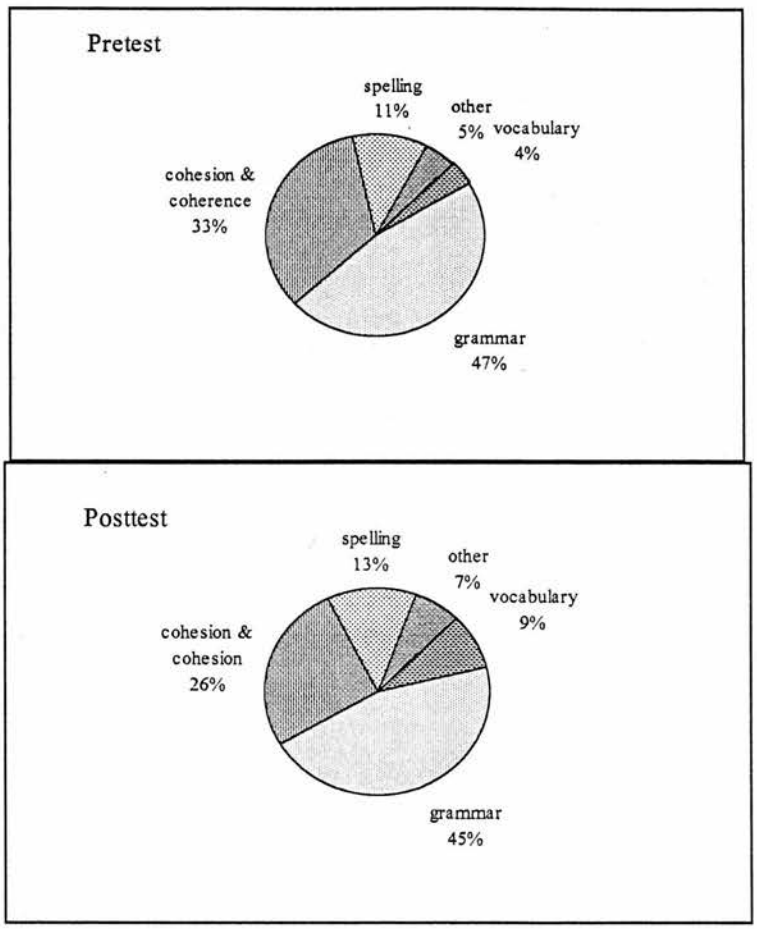

Figure 2: PHN- Proportion of types of error in the pretest and the posttest

The proportion of errors in the 'cohesion and coherence' dropped sharply while the proportion of 'vocabulary' errors increased. The proportion of errors in the 'grammar' category fell slightly while proportions of other categories remained similar. It seems as though the students were getting a grip on punctuation and beginning to write more coherently with a slightly better command of grammar, so this was giving them the confidence to become more adventurous with vocabulary. Increased risk-taking would account for the rise in vocabulary errors.

8.2.2 ISN

8.2.2.1 Grammatical structure

Once again, the number and length of t-units were compared and the results for ISN are given in Table 47 below.

Table 47: ISN- Change over time in grammatical structure

t-tests n=68	pretest mean	posttest mean	change	t	p
Grammatical Structure					
t-units (per 100 words)	8.58	8.19	-0.389	1.41	0.16
words per t-unit	12.00	13.08	+1.08	1.96	0.053
error-free t-units (per 100 words)	4.40	4.22	-0.18	-0.60	0.55
words per error-free t-unit	10.25	10.64	+0.39	0.91	0.36

In contrast to PHN, where the number of t-units became more numerous and shorter as competence developed, the t-units in the imagined stories became fewer and longer. There were no significant changes although the increase in t-unit length approached significance. A closer look at the number and length of t-units (both conventional and error-free measures) showed that the writing structures of the narrative types had become more similar as time had gone on. At the beginning of the experiment, when none of the students had experience in ISN, there were significant differences between the two narrative types. That the students wrote longer t-units by the time of the posttests could indicate an increased confidence and familiarity with ISN, a previous lack of which had caused subjects to write very short, careful sentences. (See Chapter 7.2.1.)

8.2.2.2 Fluency

There was a significant increase in the fluency measure for ISN as shown in Table 48 below.

Table 48: ISN- Change over time in fluency

t-tests n=68	pretest mean	posttest mean	change	t	p
Fluency					
av. number of words per essay	268.6	368.3	+99.7	6.72	0.0000*
*significant ($p<0.05$)					

The fluency increase in ISN was marked. The subjects wrote almost 100 words more on average for the posttest essays than they had written for the pretest essays, displaying a similar increase in fluency to that shown in PHN.

8.2.2.3 Accuracy

Change over time in overall accuracy and in types of error for ISN are summarised in Table 49 below.

Table 49: ISN- Significant change over time in accuracy measures

t-tests
measures= average no of errors per 100 words
n=68

	pretests	posttests	change	t	p
Accuracy					
Grammar:					
redundancy	0.38	0.21	-0.17	2.42	0.017*
Cohesion & Coherence:					
punctuation	1.42	0.73	-0.69	3.48	0.0008*
Total	2.76	1.80	-0.96	2.83	0.0056*
TOTAL ERRORS	8.58	7.01	-1.57	2.25	0.026*

* significant ($p<0.05$)

(See Appendix K for a complete list of changes over time on objective measures.)

The decrease in overall error between pre and posttest was significant, although the ISN level of error decreased a little less than that of PHN. Once again errors in the 'cohesion and coherence' category fell most of all. There was a significant reduction in punctuation errors. The only types of error which had been significant indicators of 'good' ISN at the time of the pretests were errors in the 'cohesion and coherence' category, so it is interesting to see that the number of mistakes with these kinds of error decreased for the whole cohort as writing competence increased. Redundancy errors also decreased significantly over time, although the overall 'grammar' category showed no marked change. Once again the interesting change was not so much in the relative number of errors overall between pre and posttest, but in the changing pattern of the kinds of error the students were making. Proportional change is given in Figure 3 below.

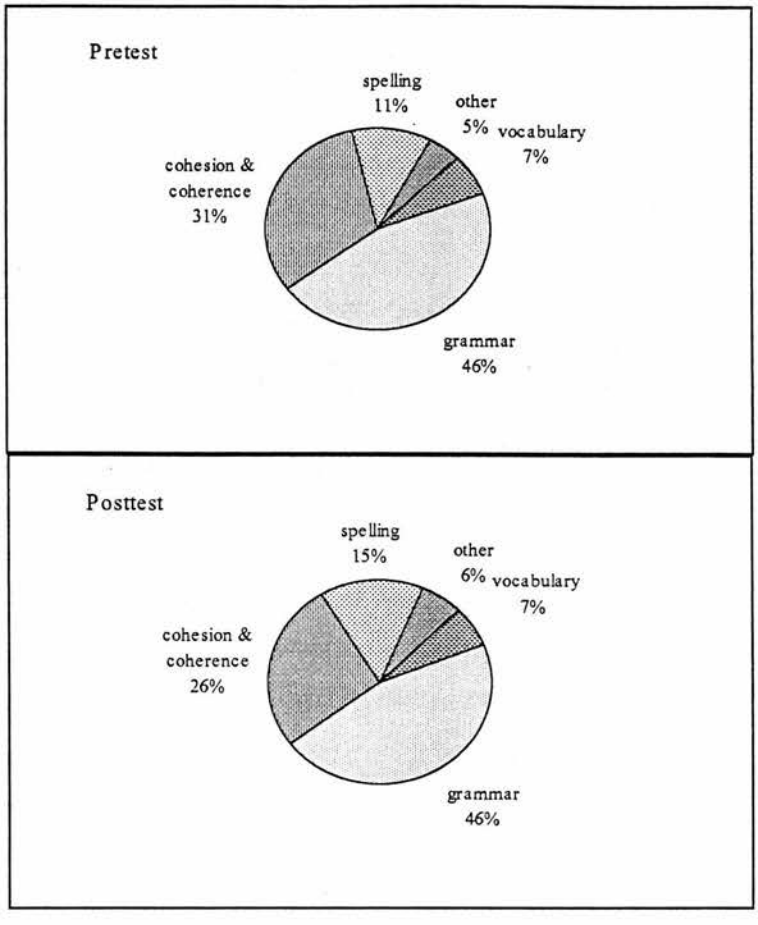

Figure 3: ISN- Proportion of types of error in the pretest and the posttest

Although the actual number of grammatical errors decreased, the proportion remained the same as it was at the time of the pretests. In contrast, the proportion of errors in the 'cohesion and coherence' category showed a substantial drop, falling from 28% at the time of the pretest to 24% at the time of the posttest. The proportion of spelling errors rose.

8.2.3 PW

8.2.3.1 Grammatical structure

The results for PW of change over time in grammatical structure are given in Table 50 below.

Table 50: PW - Change over time in grammatical structure

t-tests n=68	pretests	posttests	change	t	p
Grammatical Structure					
t-units (per 100 words)	6.18	5.79	-0.39	-1.78	0.078
words per t-unit	17.09	18.02	+0.93	1.40	0.16
error-free t-units (per 100 words)	1.54	1.77	+0.23	1.16	0.25
words per error-free t-unit	12.87	13.59	+0.72	1.03	0.31

The number of words per t-unit increased as writing competence developed, although the change was not statistically significant. This is what was expected to happen as students became more skilled at writing persuasive essays since a major feature of persuasive writing compared to narrative types is its substantially longer t-unit, which indicates lexical density.

8.2.3.2 Fluency

There was a significant rise in fluency between the pretests and the posttests, although the rise was only a little more than half that of the fluency increase on both narrative types of writing. The results are given in Table 51 below.

Table 51: PW- Change over time in fluency

t-tests n=68	pretest mean	posttest mean	change	t	p
Fluency					
av. number of words per essay	232.7	291.0	+58.3	4.36	0.0000*
*significant ($p<0.05$)					

8.2.3.3 Accuracy

Unlike the narrative types, the number of errors in PW increased as writing competence developed. Results are shown in Table 52 below.

Table 52: PW - Change over time in overall accuracy

t-tests n=68 average errors per 100 words	pretest mean	posttest mean	change	t	p
TOTAL ERRORS	10.01	10.33	+0.32	-0.42	0.67

Although the increase in the frequency of error was only slight and was not statistically significant, it is important to note because it shows that at an early level of writing development that students can improve overall and yet still not improve in their level of accuracy. 'Good' pieces of PW were associated with a relative lack of error compared to 'poor' essays (see Chapter 7.3.3) and it seems sensible to expect that as the students improved further in persuasive writing that their level of accuracy would rise, but the early improvements for most of the group involved an increase in fluency and a change in the *kinds* of error that they made, not in a reduction in error overall. See Table 52 for the significant changes in types of error.

Table 53: PW- Significant change over time in types of error

t-tests measures= average per 100 words	pretest mean	posttest mean	change	t	p
Accuracy (errors per 100 words)					
Vocabulary	0.91	1.28	+0.37	-2.15	0.033*
Cohesion & Coherence:					
reference	0.68	0.41	-0.27	2.22	0.029*
punctuation	1.68	1.12	-0.56	2.55	0.012*
Total	3.53	2.50	-1.03	2.95	0.0037*
Spelling	1.15	1.60	+0.45	-2.46	0.015*
Other	0.34	0.79	+0.45	-4.87	0.0000*
* significant ($p<0.05$)					

(Please see Appendix K for a complete list of changes over time on objective measures.)

The most marked decrease in error occurred in the 'cohesion and coherence' category, particularly in errors of reference and punctuation. Vocabulary errors, spelling errors and 'careless' mistakes in the category of 'other' errors rose. The proportional change in error types is given in Figure 4 below.

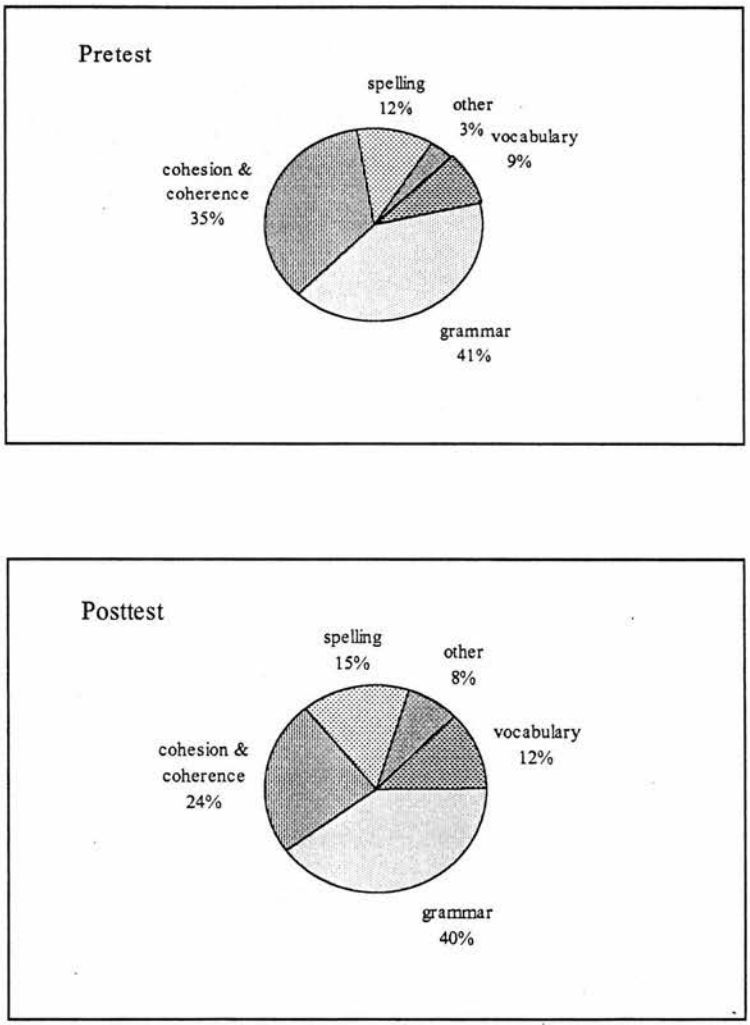

Figure 4: PW- Proportion of types of error in the pretest and the posttest

The proportion of errors in the 'cohesion and coherence' category fell from 35% to 24%. This was the most marked change between pretests and posttests. As the students' performance in persuasive writing increased, the obvious change was in their ability to manage cohesive devices better, especially reference and punctuation. The proportions of spelling and vocabulary errors and 'careless' errors (in the category of 'other' errors) rose.

8.3 Comparisons between writing types

8.3.1 Grammatical structure

The differences between writing types in the way that grammatical structure changed as writing competence developed are given in Table 54 below.

Table 54: Change over time in grammatical structure

t-tests n=68	pretest mean	posttest mean	change	t	p
Number of t-units per 100 words					
PHN	7.359	8.511	+1.152	3.68	0.0002*
ISN	8.583	8.194	-0.389	1.41	0.16
PW	6.176	5.786	-0.39	-1.78	0.078
Number of words per t-unit					
PHN	14.87	12.14	-2.73	-4.90	0.000*
ISN	12.00	13.08	+1.08	1.96	0.053
PW	17.09	18.02	+0.93	1.40	0.06

*significant ($p<0.05$)

The length of the t-units in the narrative types seemed to settle down by the time of the posttest to round about 12 or 13 words on average, which was achieved by a slight decrease in PHN and a slight rise in ISN. The much longer t-units produced for persuasive writing increased still further to an average of 18 words. The coming together of t-unit length in the narrative types by the time of the posttests to stabilise at a much shorter length than for persuasive writing is confirmed by the change shown in the error-free t-unit measure. The pattern is the same. The drawing together of t-unit lengths in the narrative types, where PHN t-units became shorter while ISN t-units became longer seems to be an evening out of writing skills in the two types of writing. At the time of the pretests, none of the students had had any practice in ISN, whereas by the time of the posttests this was no longer the case. Whatever the reason or combination of reasons for the pattern of development, it is clear that by the time of the posttests, the students were producing much longer t-units for PW than for either of the two narrative types where the t-units had developed to a roughly similar length.

It is not possible to make direct comparisons between this study and other studies of syntactic growth measured in t-units and words per t-unit, because the subjects and experimental conditions have varied so widely. It is clear though that general trends of increased lexical density associated both with syntactic growth and persuasive writing, reported elsewhere by researchers such as Hunt (1983) and Watson (1983), are shown by this study. It is not clear how the writing types would change as competence continued to develop. It is obvious that individual differences as well as differences in topic would affect the structure of the writing to some extent, but it seems too that writing types do manifest different structures, at least between narrative and non-narrative types. The average length of t-units was much shorter in the narrative types than in the persuasive writing. This is clearly demonstrated in the data and occurred even though the students were at an elementary level of competence in persuasive writing.

8.3.2 Fluency

Differences between writing types in fluency change over time are given in Table 55 and Figure 5 below.

Table 55: Change over time in fluency

t-tests n=68	pretest mean	posttest mean	change	t	p
Number of words per essay					
PHN	275.8	377.5	+101.7	6.31	0.0000*
ISN	268.6	368.3	+99.7	6.72	0.0000*
PW	232.7	290.96	+58.26	4.36	0.0000*
*significant ($p<0.05$)					

There were significant increases in fluency as measured by number of words per essay in all three writing types. There was, however, a much greater increase in fluency in the narrative types (approximately 100 words more in the posttests) than in the persuasive writing (approximately 60 words more in the posttests). This could be due to the practice the students received in narrative writing or it could have occurred because narrative is easier than persuasive writing so that fluency developed more slowly in persuasive writing. Another reason to write longer narrative than

persuasive essays is that, for most people, narrative is more enjoyable to write, so the words come out faster.

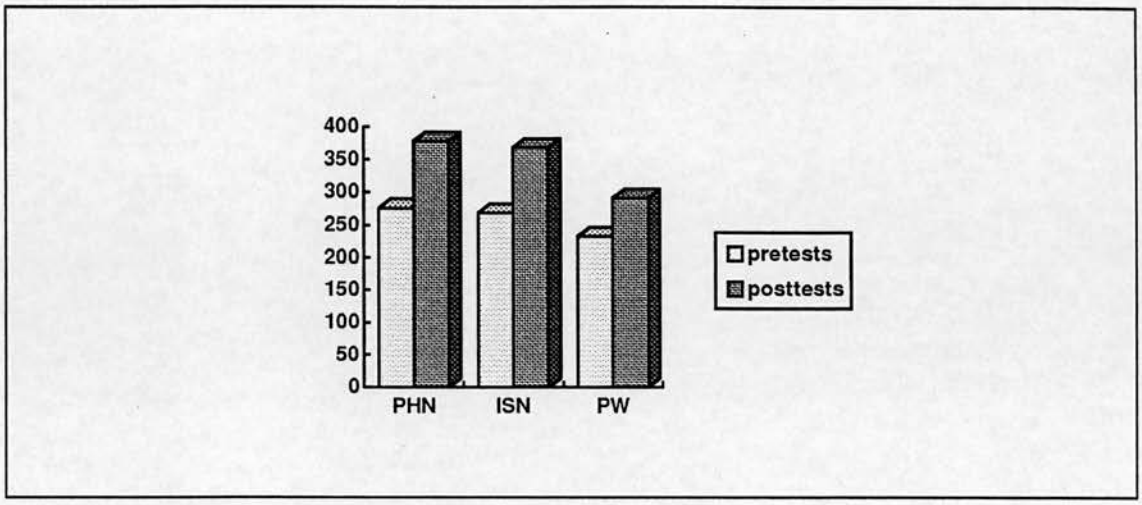

Figure 5: Increase in fluency (average number of words per essay)

Figure 5 shows the relative increases in fluency between the writing types. That increased fluency is associated with text quality and therefore with the development of writing competence is supported by previous research findings (Larsen-Freeman & Strom 1977; Larsen-Freeman 1978; Witte & Faigley 1981; Charney 1984; Homburg 1984; Ferris 1994; Kamimura 1997). The association of fluency with the development of writing competence for the subjects in this study is consistent with the findings reported in the previous chapter, where fluency was found to be a significant discriminator between 'good' and 'poor' pieces of writing in all three types. That fluency was associated with the development of writing competence was amply evidenced by the significant rise in average number of words per essay in all types of writing.

8.3.3 Accuracy

8.3.3.1 Overall error

Although overall level of accuracy was a measure which discriminated significantly between 'good' and 'poor' pieces of writing on all types (see Chapter 7), the students' writing development showed a significant decrease in error only in the narrative types and not in persuasive writing. See Table 56 and Figure 6 below.

Table 56: Change over time in total number of errors

t-tests n=68	pretest mean	posttest mean	change	t	p
Number of errors per 100 words					
PHN	8.06	6.23	-1.83	3.04	0.0029*
ISN	8.58	7.01	-1.57	2.25	0.026*
PW	10.01	10.33	+0.32	-0.42	0.67
*significant ($p<0.05$)					

The slight increase in error in persuasive writing might be because it is harder than the narrative types and the level of difficulty made progress slow. It is possible that the overall frequency of error rose slightly because students were becoming confident and taking more risks. It is possible, too, that the level of error did not decrease because of the heavier load on STM although changes in PW competence had obviously taken place as shown not only by the overall impression of improvement reported in 8.1, but also by the marked change in the fluency measure reported above. A number of points can be made in relation to observations on indicators of text quality made in the previous chapter. Firstly, ratings of quality and development can arise because a good showing on one of the objective measures may override a poor result on another. In the case of this study it seemed that fluency often seemed more important to raters than lack of error. Secondly it is a common sense observation, confirmed by the essays investigated here, that ratings are determined not solely by the objective measures on which this study focuses. This helps to explain why, despite the fact that errors did not decrease, there was still a consensus among raters that the quality of the students' persuasive writing had improved. Figure 6 below shows comparative changes in overall accuracy as competence developed.

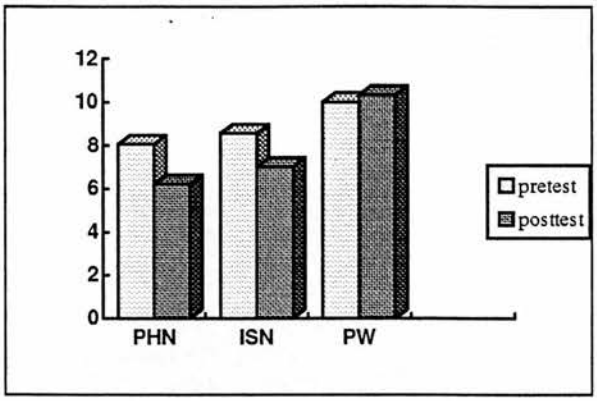

Figure 6: Change over time in overall error

Despite the fact that there is consensus that, in general, accuracy is associated with quality and therefore with the development of writing competence (Homburg 1981; Perkins 1983; Homburg 1984), previous research findings lend support to the fact there is no direct link between lack of error and quality (Evola, Mamer & Lentz 1984; Tarone, Downing, Cohen, Gillette & Murie 1993). Teachers should be aware of the possibility that an increase in error does not necessarily mean that the students have taken a step backwards. In the case of the persuasive writing produced at this level of development, it seems the writers had taken a step forwards, despite the slightly increased level of error. This finding shows that the development of competence in persuasive writing in the early stages does not necessarily have a direct relationship with the number of errors.

8.3.3.2 Type of error

The most interesting findings are, perhaps, not to do with the changes in the number of errors students made, but in the changing proportions of the kind of errors they made as writing competence developed. Table 57 below summarises the change in proportions of error.

Table 57: Summary of change over time in proportions of error

n=68 measures=errors per 100 words					
	VOC.	GRAMMAR	C & C	SPELLING	OTHER
% change	%	%	%	%	%
Increase (% rise)					
PHN	+4.67			+1.6	+2.09
ISN		+0.03		+3.58	+1.52
PW	+3.3			+3.97	+4.18
Decrease (% fall)					
PHN		-1.45	-6.92		
ISN	-0.17		-4.95		
PW		-0.39	-11.06		

(Please see Appendix E for a description of error categories and examples.)

The proportion of errors in the category of 'cohesion and coherence', i.e. errors of reference, omission and punctuation, showed a clear decrease in all writing types as performance improved. This finding supports teachers' generally perceived intuition that increased coherence is associated with increased writing performance. Interestingly, many researchers have found that errors of cohesion, such as linking words, have not been significant discriminators of increased writing competence (e.g. Mullen 1980; Tarone, Downing, Cohen, Gillette & Murie 1993)

It is difficult to make comparisons between research findings which have investigated the significance of objective measures, such as markers of cohesion and coherence, in the development of writing competence. This is because the studies have used widely differing groups, in age, proficiency, language background, circumstances, and have used different methodologies. It seems likely, however, that other studies concentrated more on linking words and less on errors of reference or punctuation. The present study did not find significant changes in the use of conjunctions, but the inclusion of punctuation in the 'coherence and cohesion' category clearly had a significant effect on the results. In PHN the drop was from 30% at the time of the pretest to 24% by the time of the posttest. In ISN, the drop was slightly less: 28% to 24%. In both types of narrative, the type of error that contributed most heavily to the change was punctuation. Errors of punctuation decreased substantially in all writing types, although punctuation received no greater attention in general essay

corrections than any other aspect of the writing. Bacha and Hanania (1980), too, in a study investigating difficulty with linking words, found that there was a marked improvement in punctuation as their subjects' writing competence developed.

In the persuasive writing essays, errors of punctuation were once again the errors which showed the biggest decrease between pre and posttests. In the pretest persuasive essays 33% of the errors belonged to the 'cohesion and coherence' category, whereas by the time of posttests, this proportion had dropped to 27%. Intaraprawat and Steffensen (1995) investigated the use of metadiscourse features (connectives, code glosses, illocutionary markers, hedges etc.) as indicators of quality in the persuasive essays of ESL university students and found that the good essays contained twice the density of metadiscourse features. Henning (1991) too found that cohesion and redundancy were significant indicators of text readability.

Since the proportion of errors in the 'cohesion and coherence' category went down, it is interesting to note which kind of errors increased. The most marked rise was in the proportion of spelling errors and this happened in all three writing types. Presumably this was due to increased risk-taking as the students' writing confidence grew. The increase in spelling errors was accompanied by a significant increase in errors of vocabulary in PHN and PW. Presumably these two phenomena, errors in spelling and errors in vocabulary, were related. Batholomae's (1980) observation that errors are difficult to classify and may arise from several causes, means that sometimes what was categorised as a spelling error may have been a vocabulary error, i.e. an actual lack of knowledge of a lexical item, and sometimes the opposite error of classification may have occurred.

In contrast to the findings here, both Charney (1984) and Grobe (1981) found that spelling errors were significant markers of text quality. Grobe, in fact, investigated ESL 5th, 8th and 11th graders, but once again it is difficult to compare findings because there are so many background variables which differ between the groups. Studies of writing development are particularly difficult to compare, since the precise span of writing development covered by a particular study is not clear.

The grade gives some indication of level, but may say more about age than about length and intensity of writing experience in the target language or in the writing type and these were not described in detail in research reports.

In summary, writing growth at this stage of development was typified by significantly increased fluency in all writing types. Frequency of error decreased significantly in the narrative types, but showed a slight increase in persuasive writing. There was a clear change in the pattern of error as competence developed, where errors in the 'cohesion and coherence' category dropped, while the proportion of spelling errors rose.

CHAPTER 9 - PRACTICE EFFECT OF IMAGINED STORY NARRATIVE ON WRITING DEVELOPMENT

The aim of the teaching given during the experiment was to compare an experimental group of students who were given practice in imagined story narrative with a control group who were given practice in personal history narrative to see if writing competence developed better in the experimental group. The PNG national curriculum did not, at the time of the experiment (although this has since changed), require more than one or two pieces of imagined story narratives from grades 1 - 10, so ISN was rarely practised and never examined. It was my belief, based on the reasons given in Chapter 3, that practice in ISN might help students to develop their writing competence in persuasive writing better than practice in PHN.

At the beginning of the experiment, before the treatment began, the students were given a pretest in three writing types: personal history narrative, imagined story narrative and persuasive writing. The scores of the control group and the experimental group were compared with a t-test to check that the groups had similar English proficiencies and there were no significant differences. (See Chapter 3.3, Table 7.) Before presenting the results of the experiment it is perhaps worth restating that instead of a comparison of one group of students who had practice in ISN with another group of students who had practice in pure PHN, what actually happened during the experiment was that each writing type was, to some extent, mixed. Although the ISN group had practice mainly in invented narrative, the control group seem to have thrown into their personal history narratives a considerable amount of invention from time to time, as it suited them.

9.1 Holistic ratings

9.1.1 Expectations

It had been expected that the ISN group would, by the end of the experiment, perform better than the control group in persuasive writing. It was hypothesised that the experimental group might perform better because some of the mental processes required for ISN seemed to be shared with persuasive writing. The following hypothesis was drawn up to test this expectation:

Hypothesis 5: Practice in writing imagined story narrative is associated with an improvement in overall performance in persuasive writing; this improvement is significantly greater than any improvement in persuasive writing associated with practice in personal history narrative.

9.1.2 Comparison between groups

The aim was to compare the development of writing competence between the groups in order to test Hypothesis 5. Gain scores between pretest and posttest on impression marks were calculated for the control group and the experimental group. An unmatched t-test was used to compare the improvement between the groups and the results are given in Table 58 below.

Table 58: Groups compared on change in holistic scores for persuasive writing

unmatched t-test to compare gain scores				
	Control Group	Exper. Group	t	p<0.05
n	(34)	(34)		
mean (of gain scores)	+1.88	+2.76		
stdev	1.17	2.12		
semean	0.20	0.36		
			-2.12	0.039*

The improvement in persuasive writing showed a significant difference between the groups. Both groups improved, but the ISN group improved by nearly 3 marks on average, while the PHN group improved by just under 2. This finding supports the hypothesis that the mental processes of imagining fostered by practice in ISN enabled the students from the experimental group to perform better in persuasive writing than their counterparts in the control group, who had received practice in personal history narrative. Hypothesis 5 was confirmed:

- Practice in imagined story narrative is associated with an improvement in overall performance in persuasive writing; this improvement is significantly greater than any improvement in persuasive writing associated with practice in personal history narrative.

Out of interest, I also compared performance between the groups in PHN and ISN. There was no significant difference between them in PHN and the experimental group performed better than the

control group in ISN. (See Appendix L for details.) These results were unsurprising and confirmed expectations firstly, that practice in ISN would enable the experimental group students to write satisfactory pieces of PHN, and secondly, that their specific practice in ISN would enable them to perform better in that type of writing than the control group. However, these results are based on trust in rater evaluations. The results in the section which follows rest on firmer ground in that they describe objective differences which are not dependent on the variability of individual judgements.

9.2 Objective measures

9.2.1 Expectations

Any significant differences in the development of structure were expected to occur in the error-free measures. This was because improvement was expected to be associated primarily with accuracy since previous research had shown that error-free measures discriminated best among holistic evaluations of ESL proficiency levels (Larsen-Freeman & Strom 1977; Larsen-Freeman 1978; Perkins 1980). Error-free measures were expected to be significant discriminators of quality so the experimental group was expected to show a greater increase in the number of words per error-free t-unit than the control group. This expectation was based on the observation that length of t-unit, as opposed to number of t-units, had been shown to be a characteristic feature of persuasive writing (Halliday & Hasan 1976). No hypothesis was drawn up to test for this expected difference, since I had decided to test the expectation of increased accuracy on the part of the experimental group through the more direct measure of total number of errors. There were no specific expectations about differences between the groups in the number or length of t-units they would produce, but only about the length of error-free t-units. However, for the sake of comparison and interest, both types of measure are reported below.

Since fluency was expected to be associated with quality and therefore with the development of writing competence, it was expected that the experimental group would write longer pieces of persuasive writing than the control group by the time of the posttests.

I decided at the time of the research design to concentrate on differences in the overall number of errors made by the two groups in the belief that this was the simplest and most straightforward way to test expectations of increased accuracy. I hypothesised that the practice in ISN given to the experimental group would increase their mental agility so that automatisation of some skills would be increased. This was expected to reduce the load on STM, which would enable students to produce more accurate writing. The following hypothesis was drawn up to test comparative improvement in accuracy:

Hypothesis 6: Practice in imagined story narrative is associated with a decrease in the number of errors in persuasive writing; this decrease in the number of errors is much greater than any decrease in number of errors associated with practice in personal history narrative.

The differences between the groups in structure, fluency and accuracy are reported in the section which follows.

9.2.2 Results

9.2.2.1 Grammatical structure

The amount of change between pretest and posttest in sentence structure, as measured by the number of words per error-free t-unit, was calculated for each student. In addition, out of interest, the other measures of structure (number of error-free t-units, number and length of t-units) were also calculated. The average for each group was worked out and the groups were compared on these measures to determine whether any structural changes, that had occurred in the writing as a result of the experiment, differed between the groups. The results are given in Table 59 below.

Table: 59 - Groups compared on change in structure of persuasive writing

	Control Group	Experimental Group	t	p<0.05
n	(34)	(34)		
	mean	mean		
words per error-free t-unit (eft)	+1.27	+1.67	-0.25	0.80
efts (per 100 words)	+0.22	+0.24	-0.05	0.96
t-units (per 100 words)	-0.48	-0.30	-0.44	0.66
words per t-unit	+0.98	+0.70	-0.25	0.80
unmatched t-test to compare gain scores				

In persuasive writing, there was no significant difference between the groups in the change in length of error-free t-unit, nor in any of the other t-unit measures. Structurally, the writing of both groups had developed similarly.

Out of interest the changes in PHN and ISN structure were compared between the groups. There were no significant differences between them in PHN, but there were two areas of difference in ISN. There was a significant difference between the groups in the number of error-free t-units written for ISN (the control group's average showed a slight increase, while the experimental group's average showed a decrease) and there was a difference which approached significance ($p=0.055$) in number of words per t-unit. The experimental group produced t-units for ISN that were nearly two words longer than those of the pretests, while the control group's t-unit length increased by only a fifth of a word. It seems that in ISN the experimental group's sentence structure, as demonstrated by increased t-unit length, was changing away from a spoken language type of writing towards a more formal written style. The control group's sentence structure was changing in the same direction, but not to such a marked degree. (See Appendix L for details.)

In summary, there were no significant differences between the groups in persuasive writing structure by the time of the posttests. This finding was contrary to the expectation that the experimental group would write more accurately and thus show a difference in error-free t-unit measures.

9.2.2.2 Fluency

In the light of the findings that fluency was associated with quality, both by previous researchers (Nold & Freeman 1977, Larsen-Freeman & Strom 1977, Larsen Freeman 1978, Grobe 1981, Homburg 1984, Ferris 1994) and from the results of this study (see Chapter 7), it was expected that the experimental group would improve in fluency more than the control group.

The pre and posttests were administered under examination conditions and an hour was given for each test. The difference in the number of words written in the pretest and posttest was calculated for each student and a mean obtained for each group. The groups were compared on this measure to determine whether any changes in fluency differed between the groups. The results are given in Table 60 below.

Table 60: Groups compared on change of fluency in persuasive writing

unmatched t-test to compare gain scores on average number of words per essay				
	Control Group	Exper. Group	t	$p<0.05$
n	(34)	(34)		
mean (of gain scores)	+46.1	+56.7	-0.54	0.59

There was no significant difference between the groups, but the results show that the experimental group increased their average persuasive essay length by 10 words more than the control group. An investigation of the group differences on PHN and ISN showed the same pattern. There were no significant differences between the groups, but the experimental group increased their word average in both writing types more than the control group (PHN: 17 words more, ISN: 22 words more). (See Appendix L for details.) The lack of significant difference in fluency increase between the groups was contrary to expectations.

9.2.2.3 Accuracy

Overall.

I hypothesised that the practice in ISN given to the experimental group would cause them to show significantly greater improvements in accuracy over the control group in persuasive writing, which imposes a heavy load on STM. Hypothesis 6 was drawn up to test the expectation that practice in imagined story narrative is associated with a significant increase in accuracy in persuasive writing.

The difference between the average number of errors per 100 words made in the pretest and the average number of errors per 100 words made in the posttest was calculated for each student and a mean obtained for each group. The groups were compared on this measure to determine whether any changes in number of errors that had occurred in persuasive writing differed between the groups. Table 61 below gives the results.

Table 61: Groups compared on change of accuracy in persuasive writing

unmatched t-test to compare gain scores on average number of errors per 100 words

	Control Group	Exper. Group	t	p<0.05
n	(34)	(34)		
	mean	mean		
	+1.80	-1.16	4.14	0.0001*

There was a significant difference in the change in level of error in persuasive writing. The experimental group, who had received practice in ISN, wrote persuasive writing posttest essays that were significantly more accurate than those produced by the control group. The experimental group achieved on average a decrease of 1. 16 errors in persuasive writing, while the control group's level of error increased by 1.8 errors per 100 words. That there was a significant difference between the groups in levels of accuracy in persuasive writing was in accord with original expectations, i.e at the time when I had not known that a mixing of writing types would occur in the treatment. In view of the substantial amount of invention the students in the control group claimed to have used in their PHN treatment practice, I found it surprising that there was such a significant difference between the

groups in the level of error in their persuasive writing. There are two possible explanations. The first is that the control group students exaggerated the amount of invention they had included in their PHN treatment essays, so that the differences in types of treatment were actually more marked than they appear to have been. The second is that the difference in writing type focus, albeit relatively small if the students' comments are reliable, made a significant difference to the experimental group's ability to produce accurate persuasive writing. It seems that the practice in ISN must have enabled an increased ease of processing for those cognitive operations shared with persuasive writing. The extra practice the experimental group had received in imagining, comparing and choosing event progressions seems to have facilitated their processing powers to make their load on STM lighter than the load experienced by the control group. Consequently the experimental group seemed to have reached a point in their development where they made fewer errors in their persuasive writing.

Hypothesis 6 which stated:

- Practice in imagined story writing is associated with a decrease in the number of errors in persuasive writing; this decrease in the number of errors is much greater than any decrease in number of errors associated with practice in personal history narrative.

was confirmed.

Comparative change in levels of accuracy were also compared for PHN and ISN, but no significant differences were found. (See Appendix L for details.) The explanation seems to be that differences between the groups became apparent only when a writing type (persuasive writing) that imposed a heavier cognitive load than narrative types was tested.

A further investigation of differences between the groups in persuasive writing was undertaken to find out which error categories and types had shown significant differences. The difference for each student on each type of error was calculated and a mean obtained for each group. The groups were

then compared on each type of error to see which error types differed significantly. The results are given in Table 62 below.

Table 62: Significant differences between groups in types of error made in persuasive writing

n	Control Group (34) mean	Exper. Group (34) mean	t	p<0.05
Cohesion & Coherence				
Reference	+0.074	-0.60	2.87	0.0055*
Punctuation	-0.13	-0.98	2.24	0.028*
Total	-0.12	-1.94	3.68	0.0005*
Other errors (carelessness, style)	0.935	-0.032	6.21	0.0000*

unmatched t-tests to compare gain scores on average number of errors per 100 words

The significant differences between the groups were in the categories of 'cohesion & coherence' and 'other' errors, those of carelessness and style. The 'cohesion and coherence' errors that the students in the experimental group reduced significantly were errors of reference and errors of punctuation. In contrast to the control group, they also reduced their frequency of error in the category of 'other' errors. These included 'careless' errors, which were defined as mistakes where the student had produced a correct version previously, errors of one word or two, for example 'intime' instead of 'in time' and errors of style. There were hardly any errors of style and the 'one word or two' type errors seemed to vary little between the groups. 'Careless' errors, however, made up the largest proportion of this category and the control group made more of this kind than the experimental group, presumably because of an inability to cope with the load on STM that the production of persuasive writing was imposing.

9.3 Summary and discussion

It is not possible to set this part of the study, which attempted to test the relative benefits of one kind of narrative writing practice against another kind, in the context of other research findings. To my knowledge, no other such study has been attempted.

The hypothesis that practice in ISN aids the transition to persuasive writing to a significantly greater degree than practice in PHN was tested first of all by carrying out t-tests on the average gain scores, according to holistic evaluations, of each group. The experimental group improved significantly more than the control group on this measure, so Hypothesis 5, which stated that practice in ISN was more helpful for improving persuasive writing than practice in PHN, was confirmed. The result, however, needs to be viewed cautiously since inter-rater reliability was not high and inter-rater reliability in general seems doubtful. The results of the post hoc re-evaluation of persuasive writing scripts reported in the next chapter were different from the ratings given for the experiment and showed that the more raters that were included, the less likely it was for consensus to be achieved. The results of the experiment show, too, that practice in both types of narrative had beneficial results, since the whole cohort showed some improvement in persuasive writing.

The second expectation was the experimental group would produce more accurate persuasive writing than the control group and they did. The experimental group's level of accuracy improved by an average decrease of about one error per 100 words, while the control group's level of error in persuasive writing *increased* (by about one and three quarter errors per 100 words). The groups were significantly different on the measure of change in overall error so Hypothesis 6, which stated that the level of accuracy in persuasive writing would be benefited by practice in ISN significantly more than by practice in PHN, was confirmed.

The types of error which showed a significant decrease on the part of the experimental group compared to the control group were reference and punctuation errors in the 'cohesion and coherence' category and 'careless' errors in the category of 'other ' errors. A comparative lack of these kinds of error can be considered to result, at least to some extent, from a greater ease of text production processing, which the experimental group is argued to have achieved through practice in ISN.

There were no significant differences between the groups in terms of PW structural change or fluency, so it is noted that any greater improvement achieved by the experimental group over the

control group was apparent only in a difference in level of accuracy. Since the investigation into objective indicators of 'good' pieces of persuasive writing showed that fluency seemed to be more important than accuracy at an early level of development, the result of no difference on the fluency measure is slightly surprising. It may be that since both groups achieved a significant increase in fluency in persuasive writing, the developmental need for a certain level of fluency had been met and the next stage of development involved an increase in accuracy. It was at this point, presumably, that the groups differed.

CHAPTER 10 - POST HOC RE-EVALUATION OF PERSUASIVE WRITING SCRIPTS

The pretest and posttest persuasive essays, which had been used to test the hypothesis of increased writing competence in the experimental group, were re-evaluated by eight new raters. This was carried out firstly to see whether persuasive essay evaluations by a different set of raters would find the same differences between the control group and the experimental group, and secondly to investigate generally whether rater evaluations were a reliable method of testing.

10.1 Reasons for taking a second look at inter-rater reliability

The relationship between writing types and the development of writing competence were investigated mainly through the results of rater judgements of text, which were used either directly or indirectly. The demonstration of the hypothesised hierarchy of difficulty between the writing types used holistic scores to plot the implicational scale and the descriptions of indicators of quality and of the development of writing competence were compiled by relating objective features to holistic scores. The hypothesised beneficial contribution of practice in imagined story writing was tested only partly through the use of holistic evaluations, but, in all, rater judgements were used to test five of the six hypotheses proposed by the study. It is clear, therefore, that inter-rater reliability is crucial for most of the test results and to accept the results one has to be reasonably sure that the same findings would emerge no matter which raters were used, or when the ratings were carried out.

The inter-rater reliability of the three raters who were used for the study was not found to be high. On the persuasive writing pretests it was .55 to .65, and on the posttests it was .39 to .55. These figures mean that on the pretests rater agreement varied from 30 to 42% (shared variance) and on the posttests from 15 to 30%. By looking only at the Pearson correlation coefficient (r) and noting that this is significant at 0.01, while not looking at the actual shared variance between raters, it is easy to overlook the level of disagreement. Warnings from an ever increasing number of researchers of problems with rater evaluations, however, should not be ignored. There can be substantial variation in rater harshness even with rater training (Charney 1984, Lumley & McNamara 1995) and rater characteristics are not always consistent over time (Lumley & McNamara 1995).

A further problem arises in that the validity of ratings may be compromised if raters are forced to conform to preset criteria (Charney 1984; Barrit, Stock & Clark 1986; Huot 1990; Henning 1991; Horowitz 1991; Wiegle 1994) as they are in the ratings for the TWE where high inter-rater reliabilities are reported (Stansfield 1986). Although the raters for this study had been asked to conform to a holistic evaluation rating scale, there was no attempt to adjust evaluations where raters had varied, so each rater's judgement was treated as valid. I discussed the rating scale with each of the raters before marking started, but the scale obviously did not have the intended effect of making it easy for raters to agree on the value of scripts. Polio (1997) found that the holistic scale she used in her study was problematic because inter-rater reliability was low, and yet the raters did not feel that the scale could be modified to make it more reliable.

For the reasons outlined above it was decided to perform a post hoc re-evaluation of the persuasive writing scripts which had been used to test Hypothesis 5:

- Practice in imagined story narrative is associated with an improvement in overall performance in persuasive writing; this improvement is significantly greater than any improvement in persuasive writing associated with practice in personal history narrative.

10.2 Description of the post hoc rating

10.2.1 Raters

Eight teachers with qualifications and experience similar to the first set of raters were asked to perform a post hoc re-evaluation of the persuasive writing essays to see if the result shown by the first set of raters would be repeated. The second set of markers consisted of five PNG NNSs (three males and two females and 3 expatriate NS raters (one male and two females). (The raters used for the experiment had been one male PNG NNS, and two female British NSs.) All the markers were experienced teachers, with similar levels of experience to those raters who evaluated for the experiment. The composition of raters for the post hoc evaluation contained more females and more PNG non-native speakers, as well as there being nearly three times as many raters as originally employed in the experiment. More raters were used deliberately in the hope that more raters would

produce more reliability of evaluations, as Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey (1981) have suggested.

The second set of raters evaluated only the persuasive writing scripts and received the pretests and the posttests at the same time. This was different from the procedure used with the ratings for the experiment, where the pretest scripts had been given to the raters immediately after the pretests, followed by a gap of several months when they received the posttest scripts. Both the raters for the experiment and the post hoc raters knew which scripts were pretests and which were posttests because they were labelled as such, but the students' names had been removed and replaced with numbers, so none of the raters in either set knew whose script they were marking. As with the first raters, the marking took place over a few weeks. The raters in the second set used the scoring guide (see Appendix D), as the first set of raters had done, to give a holistic impression mark out of 5 for each essay. The markers in each set were different people, i.e. no person rated in both sets.

10.2.2 Results

10.2.2.1 Overall improvement of whole cohort

The raters for the post hoc evaluation found a significant difference in the performance of the whole cohort between the pretest and posttest persuasive writing, but the improvement they found was not so great as that found by the raters for the experiment. See Table 63 below.

Table 63: Improvement over time on PW - post hoc re-evaluation

	pretests	posttests	change	t	p<0.05
n	68	68			
	mean (/5)	mean (/5)			
	2.465	2.793	+ 0.328	-4.22	0.0000*
cf. Experiment rating:					
	pretests	posttests	change	t	p<0.05
n	68	68			
	mean (/5)	mean (/5)			
	2.40	3.18	+0.78	-7.01	0.0000*

t-tests

The results of both post hoc and experiment ratings showed a significant improvement in persuasive writing, but the post hoc ratings were slightly higher for the pretests and lower for the posttests. This could have been due to the fact that the post hoc evaluators received both pretest and posttest scripts at the same time, unlike the raters for the experiment, and so were better able to judge differences between the pretest and the posttest essays, or it may simply have been due to individual differences in ways of evaluating.

10.2.2.2 Effect of practice in ISN

The experiment ratings had shown a significant difference ($p=0.039$) between the control group and the experimental group, where the experimental group, who had received practice in ISN, showed a greater improvement. In contrast, the results of the post hoc evaluations found no significant difference in improvement in persuasive writing between the experimental group and the control group. See Table 64 below.

Table 64: Groups compared on change in holistic scores for persuasive writing - post hoc

	Control Group	Exper. Group	t	$p<0.05$
n	(34)	(34)		
	mean	mean		
	2.65	2.21	0.52	0.61
unmatched t-test to compare gain scores				

The results from the second set of evaluations showed no significant difference between the groups although the control group improved slightly more on average than the experimental group. Unlike the results from the experiment ratings, the post hoc results would not have confirmed Hypothesis 5, which stated that practice in imagined story narrative would be associated with an improvement an overall performance in persuasive writing significantly greater than any improvement associated with practice in personal history narrative.

It is interesting in the light of the post hoc result, which found no highly significant difference in improvement between the groups, to consider the results of the comparison of accuracy levels

between the groups. The accuracy level of the experimental group was found to have improved to a significantly greater degree than that of the control group. The test for comparison of improvement in accuracy levels was carried out by counting the change in errors between pre and posttest for each student and averaging them for the groups. It was an objective test that did not rely on subjective evaluations. The rater evaluations from the experiment found that the experimental group had improved in persuasive writing significantly more than the control group, while the results from the post hoc re-evaluation found that the control group had improved slightly more than the experimental group. The difference between the two sets of evaluations could mean either that the increased accuracy on the part of the experimental group should not be interpreted as a sign of increased performance, or that rater evaluations are an unreliable way of coming to conclusions about improvement in writing competence.

Inter-rater reliability for the post hoc re-evaluation, which is reported in the next section, can be seen to be lower than it was for the experiment, but it could be argued that this does not necessarily mean that the ratings for the experiment were more valid. It seems that when more raters are used, there are more conflicts in rating. It is not possible to decide who is right and who is wrong, or whether necessarily there has to be a 'wrong' and a 'right' where there is a conflict.

10.2.3 Inter-rater reliability

Inter-rater reliability for the persuasive writing pretests ranged from .02 to .60 (shared variance range from 0 to 36%). See Table 65 below.

Table 65: Inter-rater reliability for post hoc evaluation of persuasive writing pretests

	R1	R2	R3	R4	R5	R6	R7	R8
R2	0.419*							
R3	0.294	0.322						
R4	0.279	0.207	0.071					
R5	0.328*	0.366*	0.264	0.044				
R6	0.418*	0.287	0.221	0.059	0.481*			
R7	0.353*	0.360*	0.168	0.020	0.330*	0.305		
R8	0.595*	0.401*	0.109	0.361*	0.258	0.185	0.427*	
preav	0.763*	0.707*	0.471*	0.348*	0.656*	0.648*	0.596*	0.660*

* sig at 0.01

Key: R1, R2, R3 = male PNG NNS raters
R4, R5 = female PNG NNS raters
R6 = female expatriate NS rater
R7, R8 = male expatriate NS raters
preav = pretest average

Pearson correlations (*Significant correlation is 0.3248 or above for a two tailed test at 0.01.)*

Despite the fact that all raters correlated significantly with the pretest average, only 12 out of a possible 28 pairs of raters were significantly correlated with each other. Raters 3 & 4 (male NNS/female NNS)), raters 4 & 5 (female NNS/female NNS), 4 & 6 (female NNS/male NS), 4 & 7 (female NNS/female NS) found few points of agreement with each other. It seems unlikely from a consideration of the pairs of raters whose evaluations were not significantly correlated that disagreements over evaluation could be traced to either a NS/NNS split, or to a male/female split.

On the persuasive writing posttests inter-rater reliability dropped to a range of -.14 to .44. See Table 66 below.

Table 66: Inter-rater reliability for post hoc re-evaluation of persuasive writing posttests

	R1	R2	R3	R4	R5	R6	R7	R8
R2	0.293							
R3	0.053	-0.029						
R4	0.122	0.245	-0.137					
R5	0.308	0.241	0.097	-0.015				
R6	0.357*	0.193	0.056	-0.061	0.350*			
R7	0.397*	0.424*	0.379*	0.048	0.249	0.311*		
R8	0.424*	0.365*	0.081	0.295	0.365*	0.166	0.442*	
poav	0.642*	0.606*	0.295	0.249	0.581*	0.638*	0.705*	0.657*

* sig at 0.01

Key: R1, R2, R3 = male PNG NNS raters
R4, R5 = female PNG NNS raters
R6 = female expatriate NS rater
R7, R8 = male expatriate NS raters
poav = posttest average

Pearson correlations (*Significant correlation is 0.3248 or above for a two tailed test at 0.01.)*

On the posttest ratings only 10 out of a possible 28 pairs of raters were significantly correlated with each other. Raters 3 and 4 did not correlate significantly with the posttest average. Raters 3 and 4 were both PNG NNS raters and rater 3 was male while rater 4 was female. Once again, this seems to rule out two possible sources of variability: NS versus NNS differences and male/female differences. A close look at the correlations shows that raters 1 & 3 (male NNS/male NNS), raters 3 & 6 (male NNS/male NS), raters 4 & 7 (female NNS/female NS) achieved little agreement. It is true that raters 4 & 7 showed little agreement with each other on the pretests, but raters 1 & 3 had correlated at .294 which dropped to .053 on the posttests. This indicates that raters 1 & 3 were either being inconsistent and had changed their personal criteria by the time they marked the posttests, or that the posttest essays called forth different foci of evaluation. The time difference between pre and posttest marking was reduced for these post hoc raters since they were given pre and posttests at the same time and this may have had an effect.

Several of the pairs of raters had even correlated negatively with each other: raters 2 & 3 (male NNS/male NNS), raters 3 & 4 (male NNS/female NNS), raters 4 & 5 (female NNS/female NNS), and raters 4 & 6 (female NNS & male NS). Negative correlations are the most disturbing of all because it means that the pairs of raters are marking 'in opposite directions' i.e. what one rater

considers to be good, the other considers to be bad. On the pretests, raters 2 & 3 had previously correlated positively at .322. It is clear that they no longer agreed with each other but it is not clear what the reasons were for their disagreement. It is possible that 136 essays to mark, given in one batch, were experienced by the post hoc volunteer raters as a difficult overload on top of their heavy full-time teaching commitments, so that by the time they got to the posttests (assuming they marked the pretests first) they were so tired that their judgement was different from normal. The volunteer raters for the experiment, however, had received 204 essays in each batch (68 x 3 pretests, then 68 x 3 posttests), although at least they had the variety provided by three different writing types rather than 136 persuasive essays. Maybe the post hoc raters suffered from boredom after marking so many of the same kinds of essay, and this made it difficult for them to concentrate. I was very grateful to all the raters for the time given freely and with goodwill and their professionalism is not in question. Tiredness, however, must have been unavoidable and probably played in part in both sets of ratings.

In view of the low rate of inter-rater reliability, further tests were carried out to compare NS versus NNS ratings and individual differences.

10.2.3.1 Native speaker versus non-native speaker ratings

Table 67: NS v NNS inter-rater reliability post hoc evaluation

persuasive writing pretest	persuasive writing posttest
NNS v NS averages p = 0.657*	NNS v NS averages p = 0.65*

* sig at 0.001
Pearson correlations

It seems from Table 67 above that there were no differences between the native speaker and the non-native speaker groups of raters, on either the pretest or the posttest evaluations. What happened, however, was that the group averages masked disagreements between individuals in both groups.

10.2.3.2 Individual differences

Non-native speaker raters were compared to investigate their level of agreement with each other. The results are given in Table 68 below.

Table 68: NNS within-group differences

NNSs on persuasive writing pretest					NNSs on persuasive writing posttest				
	R1	R2	R3	R4		R1	R2	R3	R4
R2	0.419*				R2	0.293			
R3	0.294	0.322			R3	0.053	-0.029		
R4	0.279	0.207	0.071		R4	0.122	0.245	-0.137	
R5	0.328	0.366*	0.264	0.044	R5	0.308	0.241	0.097	-0.015
Pearson correlations									

Of the non-native speakers, only raters 1 and 2, and raters 2 and 5 agreed. on the pretest evaluations. No pairs of NNS raters agreed on the posttests and raters 2 & 3, raters 3 & 4 and raters 4 & 5 disagreed to the point where their correlations became negative. The native speaker raters were also compared to investigate their levels of agreement. The results are shown in Table 69 below.

Table 69: NS within-group differences

NSs on persuasive writing pretest			NSs on persuasive writing posttest		
	R6	R7		R6	R7
R7	0.305		R7	0.311	
R8	0.185	0.427*	R8	0.166	0.442*
Pearson correlations					

Only one pair of NS raters (7 and 8) out of three pairs found significant agreement on the pretests and posttests.

The above findings support findings of my earlier study (Phillip 1994) that inter-rater reliability problems were traceable not to NS versus NNS variability, but were caused by individual differences. In the earlier study I compared evaluations of PNG NNS raters with expatriate NS raters who had experience of PNG and NS raters at Edinburgh University who had no experience of PNG culture.

10.3 Comparison of experiment and post hoc ratings

Firstly the two sets of evaluations will be compared and then the differences in marking for both sets of raters will be investigated.

10.3.1 Pretest and posttest scores

T-tests were carried out between the two sets of marks for both the persuasive writing pretests and the persuasive writing posttests. The results are given in Tables 70 and 71 below.

Table 70: Comparison of two sets of scores on the persuasive writing pretests

number of pretest scripts=68			
exper. raters (3) mean /5	post hoc raters (8) mean /5	t	p
2.407	2.465	-0.60	0.55

There was no significant difference between the two sets of scores on the pretest.

Table 71: Comparison of two sets of scores on the persuasive writing posttests

number of posttest scripts=68			
exper. raters (3) mean /5	post hoc raters (8) mean /5	t	p
3.182	2.793	4.12	0.0001*
*significant ($p<0.05$)			

There was a highly significant difference between the two sets of evaluations on the posttest. The raters for the experiment marked much higher than the post hoc evaluation raters. The experiment raters gave an average of 3.2 out of 5 for the posttest scripts, while the post hoc evaluation raters gave an average of only 2.8.

10.3.2 Rater differences

Inter-rater reliability was calculated through Pearson correlations for all the raters who had marked the scripts, and then Analysis of Variance was calculated to describe and comment on mark ranges.

10.3.2.1 Correlations

PW pretest correlations for all raters are given in Table 72 below.

Table 72: Correlations on persuasive writing pretests - all raters

		NNS ExR1	*NS* ExR2	*NS* ExR3	*NNS* PhR1	*NNS* PhR2	*NNS* PhR3	*NNS* PhR4	*NNS* PhR5	*NS* PhR6	*NS* PhR7
NS	ExR2	0.558*									
NS	ExR3	0.648*	0.549*								
NNS	PhR1	0.327*	0.405*	0.333*							
NNS	PhR2	0.257	0.307	0.331*	0.419*						
NNS	PhR3	0.280	0.438*	0.234	0.294	0.322					
NNS	PhR4	0.140	0.096	0.075	0.279	0.207	0.071				
NNS	PhR5	0.504*	0.339*	0.394*	0.328*	0.366*	0.264	0.044			
NS	PhR6	0.308	0.286	0.229	0.418*	0.287	0.221	0.059	0.481*		
NS	PhR7	0.402*	0.423*	0.260	0.353*	0.360*	0.168	0.020	0.330*	0.305	
NS	PhR8	0.349*	0.338*	0.393*	0.595*	0.401*	0.109	0.361	0.258	0.185	0.427*

* = significant at 0.01

Key: ExR1,2,3 = Experiment Rater 1, 2, 3 (rater 1 was male, raters 2 &3 were female)
PhR1,2,3,4,5,6,7,8 = Post Hoc Rater 1,2,3,4,5,6,7,8 (raters 1,2,3,6 were male; raters 4,5,7,8 were female)
NS = Native Speaker
NNS = Non-Native Speaker

Pearson Correlations (*Significant correlation is 0.3248 or above for a two tailed test at 0.01.)*

It is clear from Table 72 above that some markers from both sets (those raters who were used for the experiment and those raters who did the post hoc evaluations) correlated significantly with each other, while others did not. The correlations between pairs of raters ranged from .02 to .65 so shared variance ranged from 0 to 42%. The strongest agreement was achieved between experiment raters 1 & 3. There was no clear division between NS raters and NNS raters, nor between males and females, although the raters used for the experiment achieved more agreement with each other than with the post hoc evaluation raters. The differences in ratings seemed to be mainly between individuals and possibly between the two sets of raters. A difference between the two sets of raters could have been due either to conditions of marking or other background variables or to individual differences.

The PW posttest correlations for all raters were calculated and are given in Table 73 below.

Table 73: Correlations on persuasive writing posttests - all raters

		NNS ExR1	*NS* ExR2	*NS* ExR3	*NNS* PhR1	*NNS* PhR2	*NNS* PhR3	*NNS* PhR4	*NNS* PhR5	*NS* PhR6	*NS* PhR7
NS	ExR2	0.385*									
NS	ExR3	0.550*	0.475*								
NNS	PhR1	0.389*	0.209	0.391*							
NNS	PhR2	0.378*	0.122	0.457 *	0.293						
NNS	PhR3	0.118	-0.029	0.235	0.053	-0.029					
NNS	PhR4	0.115	-0.054	0.036	0.122	0.245	-0.137				
NNS	PhR5	0.407*	0.262	0.569*	0.308	0.241	0.097	-0.015			
NS	PhR6	0.326*	0.138	0.387*	0.357*	0.193	0.056	-0.061	0.350*		
NS	PhR7	0.425*	0.096	0.571*	0.397*	0.424*	0.379*	0.048	0.249	0.311	
NS	PhR8	0.405*	0.202	0.540*	0.424*	0.365*	0.081	0.295	0.365*	0.166	0.442*

* = significant at 0.01

Key: ExR1,2,3 = Experiment Rater 1, 2, 3 (rater 1 was male, raters 2 &3 were female)
PhR1,2,3,4,5,6,7,8 = Post Hoc Rater 1,2,3,4,5,6,7,8 (raters 1,2,3,6 were male; raters 4,5,7,8 were female)
NS = Native Speaker
NNS = Non-Native Speaker

Pearson Correlations *(Significant correlation is 0.3248 or above for a two tailed test at 0.01.)*

The same pattern emerges from a scrutiny of the posttest correlations. The differences seem to be individual differences, not differences between groups of NS v NNS raters, nor differences between males and females. The most disturbing feature of the posttest evaluations is that several pairs of raters correlated negatively with each other, which indicates severe disagreement on evaluations. Table 74 below shows the number of significant agreements on the posttests.

Table 74: Persuasive writing posttests - no of significant agreements per rater

rater	*no of sig. agreements out of possible 10*
Ex R1 (male NNS)	8
ExR2 (female NS)	1
ExR3 (female NS)	6
PHR1 (male NNS)	5
PHR2 (male NNS)	4
PHR3 (male NNS)	1
PHR4 (female NNS)	0
PHR5 (female NNS)	4
PHR6 (male NS)	4
PHR7 (female NS)	6
PHR8 (female NS)	6

Table 74 shows again that differences are difficult to trace to single variables such as mother tongue, gender, or the time difference between evaluations. It is tempting to conclude that Experiment Rater 1 was the 'best' because he managed to agree with 8 other raters i.e. most of them, and that Post Hoc Rater 4 was the 'worst' because she did not agree with any of the others. Such a conclusion arises because of our need to rely on consensus. Testers of writing have to rely on consensus evaluation because it is all there is. The only way such a conclusion could be tempered or changed would have been to conduct a discussion session with the raters with the aim of discovering the reasons for the individual evaluations in the hope of bringing raters closer together in their decisions. This was deliberately not carried out in this study. The results would certainly have been informative, and the overall evaluations may well have changed through discussion. Such a session was not carried out because of considerations of validity because I was concerned that NNS viewpoints and dominant personality viewpoints should not be imposed on other raters.

There are two more points to make in connection with a view that Experiment Rater 1 was the 'best' evaluator, and Post Hoc Evaluator 4 was the 'worst'. The first point is that a group of 11 raters is still a small sample, even though in writing examinations decisions are normally made on a basis of two ratings, or even one. With a much larger group of raters we may have found that Post Hoc Rater 4 agreed with far more raters than Experiment Rater 1. The second point is to ask who knows best. Who can be sure that the majority comes to the 'correct' decision? Majority decisions on all aspects of life can be seen to change on a regular basis as history progresses.

10.3.2.2 Mark ranges

The best way to compare the relationship of one rater's marks with another's is through correlations, which have been reported and discussed above. Correlations compare the level of agreement between raters on specific scripts. The ANOVA given here is considered useful to show the ranges of marks used by individual raters and how these varied. It is emphasised that even though two raters might have almost identical mark ranges and averages, their ratings may not have correlated well because

they may have made different decisions on specific scripts. The results are given in Table 75 (pretests) and Table 76 (posttests) below.

Table 75: ANOVA on all raters - persuasive writing pretests

ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	p
FACTOR	10	91.040	9.104	15.60	0.000
ERROR	737	430.029	0.583		
TOTAL	747	521.070			

Raters' Averages and Ranges of Marks on Persuasive Writing Posttests
(based on pooled stdev)

RATER	NNS/NS	N	MEAN	STDEV	2.00 2.50 3.00
ExR1	NNS	68	2.0882	0.6632	(---*--)
ExR2	NS	68	3.0000	0.8464	(---*---)
ExR3	NS	68	2.1324	0.7708	(---*--)
PhR1	NNS	68	2.5588	0.8704	(--*---)
PhR2	NNS	68	2.2500	0.9203	(---*---)
PhR3	NNS	68	2.4118	0.6043	(--*---)
PhR4	NNS	68	2.7647	0.5496	(--*---)
PhR5	NNS	68	2.8971	0.8311	(---*---)
PhR6	NS	68	1.8676	0.8622	(--*---)
PhR7	NS	68	2.7206	0.7091	(--*---)
PhR8	NS	68	2.2500	0.6775	(---*---)

POOLED STDEV = 0.7639 2.00 2.50 3.00

The differences in mark ranges appear to be individual differences rather than differences between NS and NNS, or between raters used for the experiment and raters used for the post hoc evaluation

Table 76: ANOVA on all raters - persuasive writing posttests

ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	p
FACTOR	10	135.717	13.572	22.32	0.000
ERROR	737	448.074	0.608		
TOTAL	747	583.790			

Raters' Averages and Ranges of Marks on Persuasive Writing Posttests
(based on pooled stdev)

RATER	NNS/NS	N	MEAN	STDEV	2.50 3.00 3.50
Ex-rater1	NNS	68	3.7353	0.7252	(---*--)
Ex-rater2	NS	68	3.4412	0.8531	(---*---)
Ex-rater3	NS	68	2.3676	0.8086	(--*---)
PH-rater1	NNS	68	2.6618	0.6826	(--*---)
PH-rater2	NNS	68	2.5735	0.8343	(--*---)
PH-rater3	NNS	68	2.6618	0.5070	(--*---)
PH-rater4	NNS	68	2.7941	0.5874	(---*---)
PH-rater5	NNS	68	3.2500	0.8353	(---*---)
PH-rater6	NS	68	2.3088	1.1098	(---*---)
PH-rater7	NS	68	3.0294	0.7525	(---*--)
PH-rater8	NS	68	2.9559	0.7214	(---*---)

POOLED STDEV = 0.7797

2.50 3.00 3.50

Table 76 shows that the most noticeable difference in posttest marking is that rater 1 (male NNS) and rater 2 (female NS), who evaluated for the experiment, gave higher ratings for the posttest scripts than either experiment rater 3 (female NS) or any of the post hoc evaluators. One of these raters was a native speaker and the other was a non-native speaker. This lends support to the rating patterns, which show no clear group difference between native speaker raters and non-native speaker raters. One of the raters was male and one was female, so gender difference does not seem to have played much of a part either.

The variability in the ranges of marks shown in Tables 75 and 76 above confirm the Pearson correlation findings that there were no group differences between native speaker raters and non-native speaker raters, or between males and females, even when the raters used in the experiment were included. The differences, given that the levels of experience were similar, seemed to be idiosyncratic and not easily attributable to a specific clearly visible cause. Obviously, it is easy to be lulled into a false sense of security as regards inter-rater reliability. It is usual in testing situations to

use no more than two raters, and if these raters appear to be significantly correlated and to be marking within the same range, we assume that we have achieved adequate reliability. It is often only when several raters are used that problems start to emerge. It is clear that using more raters gives more accurate information regarding the range of evaluations that experienced teachers may award, but inter-rater reliability becomes harder, not easier to achieve.

10.3.3 Effects of the rating scale

The rating scale was used by all the raters and was divided into levels of performance from 0 - 5, where 0 was 'very poor', and 5 was 'excellent'. Descriptors focussing on three areas were provided for each level. These were 'clarity & organisation', 'interest', and 'accuracy' (See Appendix D.)

It is clear that the rating scale did not achieve standardisation of ratings. The differences in mark ranges shown by the ANOVA tables in the previous section demonstrate that the levels of the rating scale were perceived quite differently by individual raters. Consider, for example, the different ranges employed by experiment raters 1 and 3 compared with experiment rater 2. The highest mark given by raters 1 and 3 was lower than the lowest mark given by rater 2 on the persuasive writing posttests. It is clear, too, from the negative correlations that occurred, that either the descriptors were perceived differently by individuals, or they were ignored, or a mixture of both. Unfortunately, no feedback was obtained on how the rating scale was used. Such information would have been useful in clarifying reasons for differences.

10.4 Summary and discussion

There was a clear difference in overall evaluation between the raters used for the experiment and those raters used for the post hoc evaluation, particularly on the ratings for the persuasive writing posttests. There could be several reasons for the differences.

1) Gender Differences

The balance of females to males in the raters for the post hoc evaluation was 4: 4. In the set of raters used for the experiment it was 2:1. I am not aware of previous research to compare male versus female raters. Santos (1988) compared raters on many variables including age and subject specialisation, but did not investigate males versus females. In the present study, however, there were no clear differences between groups of male and female raters.

2) NNS versus NS differences

The balance of NNS to NS raters was different in the post hoc evaluation. Raters for the experiment split 1NNS: 2NS. Raters for the post hoc evaluation split 5NNS: 3NS. Some researchers have found that native speakers rate differently from non-native speakers (Kaplan 1966; Kaplan 1967; James 1977; Santos 1988; Basham & Kwachka 1991; Hinkel 1994), but the details of the evaluations performed for this study show that there were no significant group differences between native speakers and non-native speakers. The findings lend support to those researchers who have found no substantial differences between native speaker and non-native speaker ratings (Hughes & Lascaratou 1982 cited in Davies 1983). The only previous research to compare PNG NNS ratings with expatriate NS ratings found no significant difference between the groups, although many individual differences (Phillip 1994).

3) Timing of evaluations

The first set of raters evaluated the pretest essays soon after they were produced and then had a gap of many months before they evaluated the posttest essays. The raters for the experiment may have been positively disposed towards the posttest scripts simply because they knew that these were the students' posttest essays which had been written after many months of practice. However, the same observation could apply to the post hoc raters but such an expectation could have been overriden by a more accurate comparison that came about because the pretest and posttest scripts were evaluated at the same time. Close scrutiny of the ratings as discussed above does not help illuminate this issue. Only two of the three experiment raters marked substantially higher on the posttests than the post

hoc evaluators, but it is not clear what caused those raters to give the high marks they gave. It has been suggested that the experiment raters may have given high marks for the posttests because they were sympathetic to my research project and wanted to show that the students had done well, but the same possibility could apply equally to the post hoc raters since they, too, were not only colleagues, but friends. It is my belief, based on previous experience of the individuals concerned, that both sets of raters evaluated the scripts honestly and professionally.

4) Other factors

Other factors which may have caused the difference in evaluations could have been due to differences in age, in health, or in levels of stress, or workload.

The point of agreement between the two sets of raters is that the persuasive writing of both groups did improve overall, and it is worth noting that neither group of students received practice in that genre. This confirms the previous findings relating to practice in personal history narrative and imagined story narrative: namely that both kinds of writing seemed to benefit development, not only in the type practised but in persuasive writing too.

The points of difference between the two sets of raters are important to note. It could be argued that since there were more evaluators involved in the second rating, as well as the fact that the second rating evaluated the scripts at the same time, that there are grounds for considering the second evaluation to be the more reliable. On the other hand, inter-rater reliability for the experiment was much higher than for the post hoc re-evaluation which included negative correlations and this could be interpreted as grounds for considering the experiment evaluation as the more reliable. Whatever the truth of the matter, the lack of reliability between the two sets of raters casts some doubt, firstly, on the finding that practice in imagined story narrative made a significant difference to performance in persuasive writing as revealed by subjective evaluations, and secondly, on the possibility of ever being able to place strong reliance on results that depend on the judgements of readers.

The evaluations of both groups of raters make clear that despite significant correlations on both pretests and posttests for the experiment raters, and significant correlations on the pretests for the post hoc raters, that there was more disagreement than agreement and that:

- no patterns of rating were easily attributable to rater characteristics such as non-native speaker status versus native speaker status;
- raters who agreed with each other on one occasion did not necessarily agree on another;
- the more raters involved in the evaluation the less certain or reliable the evaluations seemed;
- agreement between two raters cannot be taken as an indication of reliable evaluation.

PART 4: CONCLUSIONS

CHAPTER 11 - SUMMARY OF FINDINGS ON MEASUREMENT

Measurement design can be a field full of holes that are easy to fall into and hard to climb out of. Some of the most interesting insights, however, arose because of measurement problems. With hindsight, some could have been avoided. A few pre-selected objective measures to investigate text quality, for example, were obviously inadequate and yet research has to impose limitations of some kind in order to be manageable. The effect that topics had on the subjects who were required to write on them was not always easy to anticipate, and yet some of the pitfalls might have been foreseen. The most striking problems, however, were to do with the lack of inter-rater reliability for holistic evaluations. Such problems are of concern to all those who are engaged in teaching and testing essay writing of any kind, that is all academic teachers. The research design, which relied heavily on holistic evaluation, might perhaps be forgiven in this respect since we still have no better way to evaluate text. This chapter will discuss the implications of findings on the following three issues: the limitations of objective measures as indicators of text quality or writing development; the effect of topic; and difficulties with holistic evaluation.

11.1 Limitations of objective measures

Quantifiying objective measures made two assumptions:

- that the measures chosen were significant indicators of the quality of writing;
- that the degree of quantity or absence of a feature was it most significant attribute.

The first assumption was that the measures chosen for the study would be significant indicators of quality. The research design abstracted various objective features in order to discover which of these indicated quality. The choice of measures was driven not only by evidence from previous research, but by a desire to answer the kinds of questions that teachers in Papua New Guinea ask. Is it a good idea to concentrate on grammar and, if so, on which items? Is it worthwhile to emphasise accuracy, and how big a part does accuracy play in the production of 'good' texts? Which errors contribute most heavily to failure? It is easy to focus on some text features more than others, because these are more noticeable, especially for teachers of second language writers. It is also reasonable to focus on

certain features in order to make the research manageable. It is easy to forget about the effect of others.

The objective features of writing that the study investigated were fluency, structure and accuracy. Each of these features has been found by previous research to contribute substantially to text quality and therefore to the development of writing competence. These features were also chosen because they were felt to be amenable to pedagogic intervention, and the research aimed to find information that would be useful for teachers. The assumption that the measures chosen for the study would be significant indicators of the quality of the writing, was an assumption that they would be *among* the significant indicators of the quality of the writing. This is how research usually works: by investigating a small piece of something or some of its features, and then generalising the results to apply to the thing as a whole. And yet the entirety of something, the key to its essence, cannot necessarily be accessed by the scrutiny of some of its parts.

Really good texts seem to have perfect combinations of words. 'A thing of beauty is a joy forever.....' for example, from Keats' famous poem, might feel irritating because it is overquoted, but it sticks in the mind. It feels somehow perfect, unchangeable, impossible to improve. Really good novels have the same quality of perfection, even though they may be very different from each other. This raises the question of the relationship between 'satisfactory' writing, and 'excellent' writing. Is one simply more of the other? Is excellence simply a greater amount of satisfactoriness? As teachers, we would be delighted if our students achieved 'excellent' writing, but would heave a sigh of relief if we could just get them to the minimum required, 'satisfactory writing', their springboard into writing future.

In connection with the issue of 'satisfactory' as opposed to 'excellent' writing, it is interesting to consider the fact that we rarely recognise or delight in 'excellent' academic texts, in the same way that we delight in 'excellent' fictional writing or poetry. Is this because we are left with the ideas and the arguments, rather than with the episodic structure of pleasurable sounds, rhythms, images? When we quote verbatim from academic texts, we usually copy out the relevant passage, but when we quote

from poems, for example, we often quote from memory. Maybe this is because in academic writing it is the ideas and the supporting arguments that were important and we have made these our own, to be expressed in our own words. The occasions when academic texts or speeches are memorable are usually those splurges of politician's rhetoric that deliberately employ image and emotion to imprint their ideas. Such rhetorical devices are so powerful that they can have the effect of imprinting the idea while deflecting the focus of attention from the argument that lies behind the idea. Enoch Powell's 'rivers of blood' speech is an example of this. Most people would agree that his speech had damaging effects in that it encouraged racism, but its power was undeniable, and much of its power arose from the emotional images it evoked. Two points can be made here. The first is that writing types can be mixed in that they can include features that are more usually associated with other types of writing. The second is the point that has often been made before (for example, Watson 1983): different writing types do not necessarily share the same features of 'good' text.

The assumption that the measures chosen were some of the significant indicators of text quality, was called into question by the findings described in 7.3.5. The profile of a 'good' personal history narrative, a 'good' imagined story narrative or a 'good' persuasive essay, defined by the objective measures under scrutiny, was not universally generalisable. Some of the essays did not fit. It was obvious that features other than those chosen for investigation were having an influence on the ratings.

Text organisation, the importance of which is emphasised by numerous researchers (e.g. Carrell 1982; Bamberg 1983), was not investigated. It was omitted partly for reasons of manageability and partly because of evidence that the criterion of text organisation in ratings had not been a significant discriminator between good and poor texts (Witte & Faigley 1981) and that text organisation caused most disagreement amongst raters (Phillip 1994). Findings from the present study do not make clear exactly how powerful a part good text organisation plays in essay evaluations. However, all three of the essays that did not fit the normal profile of objective measures normally associated with quality in particular types of writing, did satisfy the demands of reasonable organisation of text for that type.

The two narrative essays which were studied for evidence of features that overrode problems of inaccurate language, both satisfied Cortazzi's (1994) criteria for pleasurable organisation of narrative. These were: a beginning, a middle with dynamic change and tension, and an end that provided a resolution or outcome. The persuasive essay that did not fit the profile of 'good' persuasive essays from the point of view of accuracy, also displayed a reasonable text organisation for persuasive writing since the argument was stated and reasons for it were presented to develop the argument.

Another feature of text that contributes to quality is novelty. Novelty or surprise was stated by Barritt, Stock and Clark (1986) to have a powerful effect. They commented that when raters did not agree, they would often start discussing the writer rather than the text. The novelty factor was noticeably present in the two narrative essays that were discussed (in Chapter 7.3.5). In the personal history narrative, the bride price celebration essay had involved a bridegroom who came from a distant island. The reader was invited to anticipate, along with the village people, the possible strangeness this might introduce into the ceremony, and the excitement of the unknown. In the 'good' imagined story narrative the reader was taken on the bird's flying trip making different shapes in the air, a novel experience for most of us. From a teacher's point of view, it may be that novelty weighs even more heavily than with other professional readers, or readers for pleasure. After reading through essay after essay that say more or less the same thing, boredom starts making it difficult to concentrate. Surprise and enjoyment can count for a lot.

The importance of the sound of words and sentences must be crucial since we cannot read if we have no auditory imprint of a word (Snowling 1985). Shepherd (1994) comments on the importance of sound and rhythm to the effect of a piece of writing. Cooper and Odell (1976) tried to investigate the importance of sound in the revising processes of professional writers, but found that sound revision occupied a relatively low place in the revision hierarchy: only 19%. It may be that some revisions of which we are hardly aware, are made for reasons of sound, and that these have occurred before we became conscious of the revision process, which essentially involves second draft revisions. Sound

was not an immediately noticeable feature of the essays scrutinised for lack of fit from the point of view of accuracy, but, as readers, we may be influenced in subtle ways that are not immediately apparent. In the ISN narrative we heard the bird saying 'hello' and 'good morning' to all his dearest friends. The sound of the words are memorable possibly only because the thought of a bird saying 'hello' is unusual. In the persuasive essay, we can almost hear the writer speaking, especially when we come across the capital letters that shout at us. Rhythm, too, appeared to be an important element in carrying the reader forward through the text, particularly in the personal history narrative, where the sentences seemed to fall over themselves to get out, conveying a kind of excitement. In the persuasive essay, too, the rhythm of the writer's voice came across insistently, difficult to ignore.

Another feature which the study did not investigate was the power of imagery, and yet the two narrative essays in particular both contained images which made the text memorable. The personal history narrative had emotional images of farewell and ships setting sail, the imagined story narrative had images of a bird looping the loop with joy. Shepherd (1994) notes how the power of images contributes to text quality and memorability.

One of the most noticeable features of all three essays that received good ratings, despite the fact that they were full of language errors, was the level of audience involvement they provoked. Audience involvement is clearly an important factor and is noted by numerous researchers (Bereiter & Scardamalia 1981; Martlew 1983; Watson 1983) but it is not easy to deal with in a quantitative study because you cannot count it. It is common sense that if you bore or alienate your reader, you will not get a good rating for your essay. The reverse is also true. If the writer succeeds in making the reader feel close, if the essay gains the empathy of the reader, then it seems that he or she will forgive many other weaknesses in the writing.

The second assumption on which the value of quantifying objective measures was based was that quantity of a feature equals the importance of a feature. This assumption was challenged by the finding that some essays did not fit the general profile found by the study to be typical of a particular

writing type. The study found that high levels of accuracy were associated with 'good' essays in all types, and yet a scrutiny of the essays showed that this was not always the case. In some cases, other features of the text had clearly carried more weight with the raters than the level of accuracy. The 'average' had led to a conclusion that was not quite correct.

Even in the cases where there was no clear contra-indication that the findings of the objective measures investigated were not quite correct, such as the power of fluency, there can be no firm conclusion that a certain level of fluency is always important. Common sense tells us that long is not necessarily good, although from the writer's point of view, it may feel as though it ought to be. The effort of writing, the hard work which is needed to generate pages of words, makes the writer feel almost as though some sort of reward should be given for effort alone.

It is argued, on the basis of the reasons given above, that the findings on objective measures can be useful only so long as their limitations are kept in mind. The findings yielded the kind of results that provided the information that in general essays with more words and fewer errors were evaluated as the 'good' essays. In most respects, they provided only information that was already known. It was the findings from the 'problematic' essays which did not fit the usual profile that provided new and possibly more important information. One piece of new information was that considerations of fluency seemed to be more important than those of accuracy in cases where a conflict arose. The point to make here is that although the study has gone a little further down the road towards looking at the whole combination of factors, it still investigated only two competing features in selected cases. It did not take into account the whole text.

It is not surprising that we still have so little knowledge about what makes a good text in any writing type, despite the fact that we think we 'know one when we see one.' The combinations of so many features and the relative effect of one feature on another is obviously complex, and made more difficult to analyse because of the complication of the variability of readers. It became clear from the study that considerations of audience involvement, text organisation, sound, rhythm, and imagery

were important, despite (or maybe because of) the fact that they were not included in the research design. To analyse anything meaningfully, it is necessary to refer constantly to a view of the whole. The value of choosing a few objective measures to investigate was in the comparison of those findings with the view of a whole evaluation. This was attempted in section 7.3.5 where the problematic essays were discussed. The limitations of selected objective measures need to be kept in mind when interpreting research studies that use them.

11.2 Effect of topic

Topic is the most vital factor in the success of writing production because it determines the amount of motivation the writer feels, whether he or she wants to write, or cannot bear to write. The primary concern of the topic choice was that the prompt should generate the writing function appropriate to the type of writing to be practised. I believed that the writing function associated with a particular type of writing would very largely determine and control the level of difficulty of the essay. Attention was paid to the need to control for level of familiarity and discourse type as pointed out by various research studies (e.g. Watson 1983; Hamp-Lyons 1991e), but not enough attention was paid to other features. Audience considerations, for example, were believed to be determined and controlled largely by the writing type.

Some topic problems occurred which had not been foreseen. The first two problems were determined by the individual personal response of the writer to the prompt, and were beyond the control of the prompt setter. They still, however, need to be kept in mind. These were:

- *Inclusion of invention because of personal response to topic* - The same topics were treated differently in order to fulfil personal needs or choices of the writers to the extent that invention occurred where it was not intended by the study.
- *Variability in personal response to audience requirement for topic* - Topics showed differences in the kind of audience considerations they evoked rather than required, e.g. some topics required that the subjects recount from personal experience but called forth instead a need to invent in order to impress.

Other problems arose because some features of the prompt had not been attended to in the research design. These were:

- *Variable levels of cognitive difficulty determined by topic descriptions within the same writing type* - Close analysis of cognitive demands of various topics chosen for a single writing type revealed differences between them.
- *Variable levels of audience specification in the persuasive writing prompts* - the prompts for the pretests had implicit audience specification while the prompts for the posttests had explicit audience specification.
- *The effect on motivation and performance of a positive or negative emotional response to a topic* - Topics differed in the kind of emotional response they evoked and this caused differences in motivation and performance.

The implications of these five issues will be discussed next.

11.2.1 Inclusion of invention because of personal response to topic

The same topics were treated differently in order to fulfil personal needs or choices. This meant that invention occurred where it was not intended by the study. In the PHN group, students' personal response to some titles caused them to invent experience in order to make writing about the topic easier or pleasanter. The title 'Worst thing I have done' was perceived by some students to be unpleasant because the memory made them feel guilty. The 'Handicapped friend' title appeared to be difficult because students did not have enough real experience to fall back on. The 'Mysterious place' title was perceived by many as frightening. Some other titles, such as 'Fishy story', appeared to call forth in some students experiences that were not felt to be interesting enough to write about without elaboration.

In the cases where essay titles were perceived negatively, one of the solutions was to invent content to make them manageable and enjoyable. Titles differed in the amount of invention they evoked but

every essay inspired at least some students to include invented material. Even the essay title 'My life story' apparently inspired two students to invent the narrative. The 'Handicapped friend' title called forth the greatest amount of invented narrative, but this was presumably because of a lack of personal experience which I had not foreseen. The fourteen students who claimed to have invented what they wrote for 'The first time I watched television', however, must have done so for a different reason since there was a television at the school which all the students could and did watch. One essay on this topic showed that watching TV for the first time had been a humiliating experience, where the writer had been laughed at for his lack of sophistication. If other students had had similar experiences then it is understandable that they invented a new narrative to protect their privacy.

The kind and amount of variability in personal response to essay titles, as described above, was not anticipated. In most of the cases it could not have been foreseen and therefore lay outside the prompt setter's control. Such variability affects motivation and production, as well as imposing differing mental processes such as the inclusion of invention in PHN essays by some students and not others. Since this variability cannot be controlled by the task setter, it seems important to include a choice of task so that the student can be evaluated on his or her best writing.

11.2.2 Variability in response to audience requirement

Topics showed differences in the kind of audience considerations they evoked. The problem where students thought that the audience would require something more impressive than their own personal experience, occurred more in some essays than in others. The result was that the students invented what they considered suitable experience in order to impress. Many students invented stories for the 'Best present I ever received', for example. With hindsight it seems understandable that the students would want to impress the teacher. That might always be the case to an extent. Such a reaction had, however, not been anticipated, and the additional invention caused by this perception of myself as audience, despite the fact that it felt like a reflection of failure on my part, has to be taken into account. More importantly, it has to be concluded that audience perception is not always predictable,

it is subjective, and it is powerful in the way it determines not only style, which is much documented, but content, too.

11.2.3 Variable levels of cognitive difficulty within the same writing type

Close analysis of cognitive demands of various topics chosen for a single writing type, revealed differences between them. The intention behind the design of the essay titles used both in the experiment and for the pre and posttests had been to control the writing function in order to produce writing of one of three distinct kinds: PHN, ISN or PW. Further differences in mental processes within writing types had not been considered. The ISN titles were the ones that turned out to have problems with respect to differences in mental requirements that were actually dictated by the essay prompts. There was no performance data to show differences between pre and posttest titles because of the effect of the practice received during the writing project. It is clear, however, that there were differences in what the students were required to do. The pretest titles required the students to pretend that they were a bird, fish or pig whereas the posttest titles required the students to imagine strange experiences, where they were still human beings. At first glance this would seem to mean that the pretest titles were more difficult to write about than the posttest titles. Would it not be harder to imagine yourself as a bird, than as yourself? On further reflection, however, it might not have been more difficult because birds, fish and pigs were familiar to the students, whereas talking dogs, strange presents with knobs on and the imagination of royal lives required by the posttest titles, were not. Since there is no way of making performance comparisons on these titles because of the intervening variable of practice received during the project, we cannot be sure whether or how much the differing title requirements affected performance.

There was a further difference in ISN titles: one of the three titles required the students to imagine themselves as kings or queens, while the other two titles required students only to imagine strange events. They had to imagine meeting a talking dog or receiving an unusual present with silver knobs on, but they could still do this from the point of view of themselves. It would seem that the requirement to imagine themselves as a royal personage and then to imagine their actions and

feelings in such a role would be more difficult than imagining themselves in strange situations. However, the data for the posttests showed no significant difference in performance between the three titles.

Horowitz (1991) argues that differences between writing tasks make them into different academic genres and concludes that only very small-scale writing assessments can be valid because of this. It is not clear from the findings of this study whether different academic genres were created by different tasks or whether they were not. Neither is it clear to what extent the mental processes required by the various ISN essay prompts differed or how this interacted with other variables to affect performance. What is clear is that fulfilling the major requirement, i.e. that the prompt elicits the intended type of writing, can divert attention from other differences in task requirements which could cause differences in the level of difficulty. Attention should be paid to this issue when preparing essay prompts.

11.2.4 Implicit versus explicit audience specification

The persuasive writing pretests had implicit audience specification, while the persuasive writing posttests had an explicit specification where students were asked to write for readers of a national newspaper. The precise effect of such a difference on performance cannot be known because of the intervening months of writing practice, but a scrutiny of the scripts appeared to show both a greater variety in forms of audience address, as well as a more frequent switching between these forms for the pretests (for discussion see 6.2.3.3). This may have been caused by the difference in audience specification, or by lack of experience in persuasive writing, or by a mixture of both. It might be the case that the specific audience specification in the posttests made the writing easier for the students, but this cannot be ascertained with certainty. It is clear with hindsight that the difference between an implicit and an explicit audience specification may make a difference to performance and careful attention should be given to these matters when designing essay prompts for research or testing.

11.2.5 Performance effect of positive or negative emotional response to topic

Topics differed in the kind of emotional response they evoked and this seemed to cause noticeable differences in performance. Data from both the questionnaire responses to the treatment titles, and from performance tests on the pre and posttest titles testified to the powerful effect of the student's emotional response to the title. In the questionnaire data, the most common reason given for disliking particular titles for both PHN and for ISN was because the student disliked the experience that he or she was required to think and write about. The dislike could arise because of feelings of fear, which some students felt when required to write about ' A mysterious place' or 'Buried alive'. It could arise because of feelings of guilt, which were called forth in some students by the titles 'Worst thing I ever did' and 'Robbing a bank'. Any title which called forth a negative emotional response was perceived as being difficult to write.

In a comparison of performance on different titles, it seemed clear that differences, although not significant, depended mainly on whether a title evoked a negative or a positive emotional response. For example, the PHN posttest essay prompt the 'Worst punishment' title produced a lower average performance than the titles 'Best present' and 'First schoolfriend' (see Chapter 6.2.3.1). In ISN pretest title performance, students wrote less well about 'A day in the life of a pig' than about being birds or fishes. It was clear from an examination of their essays that it had been an unpleasant experience to imagine being a pig since the students considered them dirty and ugly (see Chapter 6.2.3.2). There was a difference, too, in performance on the persuasive writing posttest titles. The topic about the right to choose a marriage partner produced better essays than the topic about settlement in urban areas, which in turn produced better essays than the topic about road safety laws. The persuasive writing topics were different from the narrative titles in that it was the level of interest that a topic held for the student that was the positively motivating factor, rather than the level of pleasure and enjoyment that the experience evoked. The topic of the right to choose a marriage partner was dearer to the hearts of the students than the topic of road safety laws. Clearly a strong emotional response to the essay title provided the driving force for the writing.

The kind of topic that is given is a crucial factor in motivating, or not, the student to write because writing is personal and creative as well as communicative (Hamp-Lyons 1991a). From the findings of the study it seems that the students' emotional response to the topic is one of the most important factors in determining performance. Differences in performance on pretest and posttest titles were not significant but they were noticeable and they all appeared to have been caused by the difference in the students' emotional response to that topic compared with the other topics. The importance of the emotional response to the title was not appreciated at the time of the research design so care was not taken with this factor, yet it is clear with hindsight that the emotional effect is important to take into account when designing test prompts. It is interesting to note in connection with this that for a long time it has been unfashionable to make more than passing reference to the influence of affect, e.g. Kroll (1998) in her review of 'assessing writing abilities' hardly mentioned the issue. This is presumably because the role of affect has been little understood. Luria (1973), for example, in his classic text *The Working Brain*, noted specifically that the influence of affect was not yet understood. Schumann (1997), however, in a recent book on the neurological basis of affect, suggests that it may be the key factor in language learning. It seems that the role of affect in production of language is due to become the new focus of attention. Findings from this study would support such a direction.

11.3 Holistic evaluation

Although holistic evaluations are generally recognised to be the most valid form of writing assessment (Perkins 1983), inter-rater reliability is recognised to be a problem because raters react to writing in individual ways (Purves 1992). It is difficult to set criteria for evaluation, since there is no accepted theory of what makes a good piece of writing. When criteria *are* set and enforced it can be argued that validity is compromised (Charney 1984; Huot 1990; Henning 1991; Horowitz 1991). Generally we hope for the best and like to believe that when two raters agree, we are being fair to our students when we evaluate their writing. Findings from this study cast serious doubt on the possibility of writing evaluation being reliable, at least in the PNG context investigated and probably elsewhere too, since there is no reason to suppose that PNG student writing is inherently less amenable to consensus of evaluation than other second language writing.

The evaluation procedure for the research was drawn up to take into account the fact that inter-rater reliability is thought to be increased by a) an evaluation scale (Jacobs, Zinkgraf, Wormuth, Hartfiel & Hughey 1981), and b) multiple ratings (Jacobs, Zinkgraf, Wormuth, Hartfiel & Hughey 1981; Lumley & McNamara 1995). Neither of these devices succeeded in achieving inter-rater reliability, although at times they appeared to be doing so.

11.3.1 Rating scale

Most holistic ratings scales seem very general in their descriptions but this is perhaps necessary in order to allow raters to evaluate in a holistic way. The scale for the study was kept deliberately general by the overall categories of 'excellent', 'good', 'average', 'below average' or 'poor', but standardisation was sought by giving further detail in the descriptors for each level. These were divided into 'organisation & clarity', 'interest' and 'accuracy' (see Appendix D). These were the areas that research had identified as key areas of importance but it is possible that raters' attention was artificially focussed on this series of considerations and that such a focus detracted from their ability to rate holistically. Polio (1997) commented that the holistic scale she used in her study was a problem because inter-rater reliability was low, and yet the raters felt that the scale could not be modified in any way to make it more reliable. The aim of her study was to measure linguistic accuracy, and even with such a narrow focus, she felt it was not possible to construct a scale that would adequately distinguish levels of accuracy.

Data on inter-rater reliability, particularly from the post hoc evaluation, show that the rating scale did not achieve standardisation. Since some pairs of raters correlated negatively with each other, it is clear that raters' perceptions of ratings criteria were different. Since some pairs of raters marked in substantially different ranges, it is clear that the intended normative overall headings, describing level, were interpreted differently. The scale was an attempt to standardise the rating procedure but since it was not monitored, its precise effect cannot be known. It seems that rating scales should be monitored by inviting discussion among raters on how scales are used. The danger is that false

adjustment of ratings can be provoked by inter-rater relationships but information on the use of rating scales should nevertheless be included as a matter of course whenever a scale is used.

11.3.2 Multiple ratings

The received wisdom is that the more ratings that are carried out, the greater the chance there is of the ratings being reliable. The findings from this study supported my earlier findings (Phillip 1994) that the more raters employed, the more confusing the verdicts on the essays become. What can appear reliable when only two raters are employed becomes questionable when other raters evaluate the same writing and come to very different conclusions. Eleven raters in all evaluated the persuasive writing posttests (three raters for the experiment and eight raters for the post hoc evaluation). Only four of these achieved significant agreement with over half the other raters. Only one rater achieved agreement with over three quarters of the other raters. It is usually too expensive to employ more than two raters to evaluate a piece of writing, and often it is too expensive to employ more than one. This is a constraint that has to be accepted. It is very important to be aware that other raters, who are equally well qualified and experienced may come to very different conclusions.

11.4 Implications

Investigation of selected objective measures should take into account the fact that without all aspects of writing, there can be no general conclusions on what constitutes quality of text, or development of writing. We are still far from understanding how one feature of text impinges on others, and it is only when we can find a way of identifying all these and incorporating them into a single investigative design that we shall come closer to finding answers.

Designs to investigate writing should take into account the effect of the essay prompt on the subjects. The emotional effect of the topic content seems to be an important aspect of the test prompt since this determines the amount of motivation the writer brings to the task. Since there can be a wide variety of response to topics, a choice should be offered.

Our belief in inter-rater reliability has to be questioned since it has been based mostly either on evidence from just one pair of raters who have agreed on a set of evaluations, or on pairs of raters who have achieved consensus because of pressure to evaluate in certain ways. ESL professionals need to be aware of this, both for the purpose of research into writing and especially for the purpose of testing students and making decisions that will affect their lives. There are a number of researchers, who make the same point (Kaczmarek 1980; Lumley & McNamara 1995; Polio 1997), but the 'Big Tests' like TWE and IELTS continue to rely on dubious procedures of inter-rater reliability and performance on these tests is used for admission into (or exclusion from) higher education all over the world. University admissions officers should be aware of the limitations of such tests when they make their offers. Researchers should beware of conclusions based on holistic evaluations.

Holistic evaluations need to be retained since we have no better tool for investigating writing but the differences between individuals need to be recognised and accepted as valid. More research needs to be undertaken into why raters disagree both with each other and with themselves on different occasions, without starting from the presumption of some tempting perception of a platonic perfection of 'correct' evaluation (our own of course).

CHAPTER 12 - CONCLUSIONS

The main aim of the study was to track the development of writing competence in grade nine PNG high school students, to see how they managed the transition from narrative to persuasive writing, and to see whether one kind of narrative practice rather than another kind would make the transition easier. This involved first of all an investigation of the relationship between the writing types used in the study, secondly a description of how writing competence developed in the three writing types, and finally through the practice given throughout the experiment, it involved a monitoring of pedagogy aimed at enabling the development of writing competence. This chapter will summarise the findings in each of these areas and state their implications.

12.1 Relationship between writing types

It is important to emphasise that the relationship that was investigated between the writing types, was the relationship of those produced by writers at a fairly early stage of writing development. The subjects were beginners in imagined story narrative and persuasive writing. On some occasions posttest data were used for comparison to see if the relationship had changed between the writing types and sometimes it had. This means that the relationship between the types found in the study may be different from the relationship between the types of writing at a later stage of development. The findings are summarised in Table 77 below.

Table 77: Summary of findings on the relationship between writing types (pretest data)

- A hierarchy of difficulty where persuasive writing was more difficult than ISN which, in turn, was more difficult than PHN, was confirmed. (Hypotheses 1 - 4 confirmed by Gutman scaling coefficient of scalability 0.931*)

- Relationships of structure, fluency and accuracy
  Structure differed significantly between the types - ISN had the most numerous and shortest t-units, PHN fewer and longer t-units, PW had the fewest and longest t-units.

  Fluency differed significantly between PW and the narrative types, but not between the narrative types themselves. Shorter essays were written for PW.

  Accuracy differed significantly between PW and the narrative types, but not between the narrative types themselves except in errors of vocabulary and reference. ISN had more vocabulary errors but fewer reference errors than PHN.

- Indicators of 'good' scripts:

	structure	*fluency*	*accuracy*
PHN	no of error-free t-units	fluency	overall accuracy incl. 'grammar';
ISN		fluency	overall accuracy incl. 'c & c'
PW	no of error-free t-units	fluency	overall accuracy incl. 'grammar', 'c & c' & 'vocabulary'

The hypothesised hierarchy of difficulty where PW was more difficulty than ISN, which was, in turn, more difficult than PHN, was confirmed. It was no surprise that persuasive writing turned out to be harder than the narrative types, but the fact that students appeared to perform better when writing PHN than when writing ISN had to be scrutinised closely since both personal histories and imagined stories share the same narrative structure and are usually regarded as a single writing type. An examination of the implicational scale showed that the hierarchy was more stable between persuasive writing and the narrative types than between the narrative types themselves. The Gutman Scale, however, showed the hierarchy to exist, and the greater level of difficulty involved in ISN compared with PHN was supported by the students' unsolicited perceptions on the issue. The hypothesis had argued that ISN would be harder than PHN because invention requires the more stressful cognitive processes of choice and imagination.

Three points need to be made in regard to the hierarchy of difficulty. The first is that the hierarchy rests on evaluations by three raters and that other raters may have evaluated differently (as the post hoc raters did for persuasive writing). The second point is that factors other than cognitive difficulty

can have powerful effects, which can change not only the perception of difficulty but also the level of performance, for example, the emotional effect of the topic content. The third point is that the writing types were mixed in any case, and presumably what we are looking at is a hierarchy between persuasive writing with some narrative inclusions versus imagined story writing which contained some other narrative types, versus personal histories which contained some invention. The hierarchy of difficulty, where persuasive writing is harder than imagined story writing, which is in turn more difficult than the production of personal history narratives was shown to exist in this study, but for the reasons just given, such a hierarchy cannot be automatically assumed. The students' perceptions lent weight to the findings, but more research is needed in this area.

Grammatical structure differed significantly between the types. Both t-unit and error-free t-unit measures showed broadly the same pattern and, as expected, persuasive writing generated the longest t-units. Somewhat surprisingly much longer t-units were written for PHN, than for ISN at the time of the pretests. The reason for this seemed to be the writers' lack of familiarity with imagined story writing. It was argued that writers were so unused to writing invented stories that they wrote short, careful stretches of text. By the end of the experiment, however, the average t-unit length for the narrative types had come very close together, so the three writing types seemed to have settled down to conform to what had been expected: that persuasive writing would generate significantly longer t-units than the two narrative types, which would be similar. It is not surprising, but worth noting, that subjects seem to have coped with the problem of the unfamiliar writing task of imagining stories by being careful and producing short t-units. The relationship of grammatical structure between writing types, therefore, varied according to the level of writing development.

In fluency there was a significant difference only between persuasive writing and the narrative types, where PW was significantly shorter. This means that although ISN was shown by the Gutman implicational scale, as well as by student perceptions, to be harder than PHN, the increased level of difficulty was not reflected in a statistically significant way in the production of fewer words for imagined stories compared with personal histories.

Overall error differed significantly between PW and the narrative types but not between the narrative types themselves. The highest frequency of error occurred in PW, as expected. Of the two narrative types, students made more errors in PHN than in ISN at the time of the pretests. It might have been expected that ISN would generate more errors since it was shown to be harder by the implicational scale and perceived to be harder by the students. This was not the case. It seems that the students coped with the greater level of difficulty which appeared to have been involved in ISN by being extremely careful in what they wrote.

The average number of vocabulary errors was shown to differ significantly between each writing type, and in this case PHN had least errors, ISN more and PW, the most. This might mean that a greater level of difficulty associated with ISN was reflected in vocabulary errors, even though it was not reflected in the number of errors overall. However, errors of reference also differed significantly between the types, but in this case PHN contained significantly more than ISN which is not easy to explain. Persuasive writing differed from both narrative types in that it contained significantly more 'cohesion and coherence' errors. This seems understandable in the light of the greater cognitive difficulty imposed by this writing type.

That there was no significant difference in overall error between the narrative types does not conform to expectations generated by the hierarchy of difficulty between ISN and PHN, but confirms research which shows that there is only an indirect link between error and level of performance (Larsen-Freeman 1978, Perkins 1983, Homburg 1984). It seems likely, too, that the lack of correspondence between frequency of error and the hierarchy of difficulty between ISN and PHN may be due to the stages of development reached in these types of writing. As already mentioned, it seems that students coped with the greater difficulty imposed by ISN by writing careful, relatively error-free prose.

Objective indicators of quality were more similar than different for the three writing types. As far as grammatical structure was concerned, in PHN and PW the number (but not the length) of error-free t-units was an indicator of a 'good' essay, demonstrating presumably that writing development in

PW had not yet reached a stage where the length of t-unit discriminated between 'good' and 'poor' scripts, as might have been expected from a writing type characterised by lexical density. Fluency and overall accuracy were indicators of quality in all three types. The types of writing differed mainly in the *kinds* of error that discriminated between 'good' and 'poor' scripts. 'Good' PHN essays at this stage were indicated by a relative lack of grammatical error. 'Good' ISN essays, on the other hand, showed a relative lack of 'coherence and cohesion' errors. 'Good' PW scripts were characterised by a lack of error in three categories: vocabulary, grammar, and 'cohesion and coherence'. That the writers of 'poor' ISN essays were having marked difficulty with errors such as reference, omission and punctuation (the cohesion and coherence category) while the writers of the 'poor' PHN essays were not, is another indication of the different levels of competence in the two types and offers further evidence of a difference between them. It shows that despite the lesser incidence of error overall in ISN compared with PHN, the load on STM was different and imposed different cognitive constraints. In persuasive writing, the difference between 'good' and 'poor' essays was marked by more features than in either type of narrative. Poor writers struggled and made significantly more vocabulary errors, more omission, reference and punctuation errors as well as more grammatical errors than the good writers. The marked increase in the number of error types that discriminated significantly between 'good' and 'poor' PW scripts testifies yet again to the heavy load on STM imposed by this type of writing.

Findings are weakened by small sample sizes, by reliance on subjective ratings to categorise the essays, and by the fact that text features other than those under investigation clearly made a difference in some cases. An investigation of essays which did not fit the average profile of a 'good' essay in a particular type showed that considerations of fluency seemed more important than considerations of accuracy. It seems that it is crucial for a piece of writing to satisfy some kind of minimum fluency requirement in order for the writing to be detailed enough to be interesting. It would seem reasonable to expect, too, that some kind of minimum level of accuracy would be necessary, but it was clear from an examination of the essays which did not fit, that readers can forgive large amounts of error if the writing holds their interest. It became clear, too, that text

features which had not been investigated such as text organisation, sound, rhythm and imagery have powerful effects on text quality.

12.2 Development of writing competence

The whole cohort was compared to see how their writing in the three types had changed between pretests and posttests. The treatment time was approximately three terms (out of a four-term academic year) and subjects received practice in either PHN or ISN during this period. The findings are summarised in Table 78 below.

Table 78: Summary of findings on the development of writing competence

- A significant improvement in all writing types was shown by t-tests to compare pretest and posttest means (holistic ratings).
- Changes in structure, fluency and accuracy (t-tests to compare pretest and posttest means of t-unit & eft measures, average number of words per essay & number of errors per 100 words) :

  In PHN the number of t-units increased but their length decreased. Fluency and overall accuracy both increased significantly. 'Cohesion & coherence' errors showed a significant decrease, while 'vocabulary' errors rose.

  In ISN there was no significant change in structure although the length of t-units increased noticeably. Fluency and overall accuracy increased significantly. 'Cohesion and coherence' errors showed a significant decrease.

  In PW there was no change in structure. Fluency increased significantly, but overall accuracy showed a slight decrease. 'Cohesion and coherence' errors fell significantly but vocabulary and spelling errors as well as 'other' errors, e.g. carelessness, showed a significant increase.

We expect that if students are given practice in writing that their competence will develop. We do not necessarily expect that if students are given practice in one kind of writing, this will help them to improve in a different kind of writing but this is what the data suggest. The whole cohort showed a significant improvement in all three types of writing, no matter which kind of writing practice they had received. It is important to bear in mind that the findings on improvement in performance rest on subjective evaluations, and that even apparently high inter-rater reliability can mask rater differences and problems, as discussed in Chapter 10. The improvement in overall performance, however, was highly significant for all three writing types, and the post hoc re-evaluation of

persuasive writing scripts confirmed the improvement for that writing type. This leads to the conclusion that practice in PHN and in ISN helped the students' writing to improve in both types of narrative as well as in persuasive writing.

Changes over time in the structure of the writing varied between the types. Persuasive writing, for example, showed no noticeable change in structure as competence developed but the structure of PHN and ISN changed in opposite directions. The length of PHN t-units decreased significantly while those of ISN increased to a level approaching significance ($p=0.053$). This means that by the time of the posttests the number and length of t-units for the narrative types were similar, as discussed above, but at the time of the pretests when ISN was a new type of writing for the students, more and shorter t-units were produced than for PHN. For all three types of writing, the error-free t-units showed the same patterns, if not quite the same degree, as the t-unit measures which did not take error into account. This suggests that the presence of error is to be expected at an early point in writing competence and that error-free measures are not the only measures that discriminate to show improvement at this stage. This is in contrast to Perkins' (1980) finding that only error-free measures discriminated significantly to show text quality when correlated with holistic evaluations.

Fluency, as measured by the average number of words written in an hour, increased significantly for each of the writing types. The average length of PHN and ISN essays increased by approximately a hundred words, which means it was about as third as long again by the time of the posttests. The persuasive essays, which had been shorter in the first place, increased by an average of about 60 words, so that they were about a quarter as long again by the time of the posttests. Fluency could have been affected by the topic content, or by the state of the writers, but it seems unlikely that these factors could have caused such a marked increase. The topic effect was controlled to some extent by the fact that three titles were used for each writing type. There was no way to control for the state of the writers, but no general calamities occurred at the time of the tests. Two points need to be made. The first is that the length of the persuasive essays as opposed to the narrative writing was shorter at both stages of development. As competence developed, the difference in increase was less for the

persuasive writing than for the narrative writing. The second point is the obvious one that development in writing competence at the grade nine level in Papua New Guinea was marked by a highly significant increase in fluency.

Change over time in level of accuracy differed between persuasive writing and the narrative types. In the narrative types, the essays became more accurate as competence developed, but at this early stage of development in persuasive writing the average level of accuracy did not improve. In fact, at the time of the posttests the persuasive essays contained on average slightly more errors than at the time of the pretests. As writing competence developed, however, there was a change in frequency of the *kinds* of error that students made and these differed between the writing types. In PHN, the number of 'grammar' errors (articles and redundancy) and 'coherence and cohesion' errors (reference and punctuation) dropped as competence developed, while the number of vocabulary errors rose. ISN showed a similar pattern where 'grammar' errors (redundancy) and 'cohesion and coherence' errors (punctuation) fell significantly, although there was no significant rise in vocabulary errors. In persuasive writing, just as in both narrative types, the number of cohesion and coherence errors (reference and punctuation) fell significantly, but unlike the narrative types, not only the number of vocabulary errors, but also spelling errors and 'other' errors, i.e. mainly careless errors, rose significantly.

It was especially interesting to see how the *proportions* of error changed. For all three types of writing the proportion of spelling and 'other' errors increased, while the proportion of 'cohesion and coherence' errors decreased. A significant decrease in punctuation errors contributed to the fall in the proportion of the 'cohesion and coherence' category. Although the proportions of these categories of error changed in the same direction for all the writing types, they differed in degree. The proportion of spelling errors rose much more for ISN and PW than for PHN and the proportion of errors in the 'cohesion and coherence' category fell almost twice as much for PW as for either of the narrative types. More research needs to be carried out in order to show whether the findings of this study apply to the development of writing competence in ESL students generally.

12.3 Effect of practice in ISN

The study's comparison of the effect of practice in imagined story narrative with practice in personal history narrative had three motivating factors. The first was the need to find a way of helping second language students with their difficulty in writing persuasive academic essays. The second was the awareness that invented writing is practised as a matter of course in first language environments but not necessarily in second or foreign language learning situations.[1] The third was the speculation that practice in ISN appeared to share some mental processes in common with the production of persuasive writing. It seemed worth a try to see if practice in invented story writing would benefit students in the transition to persuasive writing. The experimental group received practice over several months in ISN, while the control group received practice in writing PHN. The findings are summarised in Table 79 below.

Table 79: Summary of findings on the effect of practice in ISN on the transition to persuasive writing

- The experimental group achieved a greater overall improvement in persuasive writing than the control group according to the holistic ratings (Hypothesis 5 confirmed by unmatched t-test to compare gain scores p=0.039*)

- Differences between groups in structure, fluency and accuracy in persuasive writing (unmatched t-tests to compare gain scores)

Structure – no significant difference between the groups

Fluency – no significant difference between the groups

Accuracy – level of error decreased in experimental group and increased in control group (Hypothesis 6 confirmed by an unmatched t-test to compare average change in number of errors between groups p=0.0001*)

According to a comparison of gain scores on the holistic ratings, the experimental group had improved in persuasive writing significantly more than the control group by the end of the project, so Hypothesis 5, which stated that practice in ISN would be associated with a significantly greater improvement in persuasive writing than practice in PHN, was confirmed. However, a post hoc re-

[1] It was not practised in mainstream secondary education in Papua New Guinea in 1990 when the study took place, but in the mid nineties invented stories were included in curriculum plans for the new genre-based writing syllabus.

evaluation of the persuasive writing scripts showed no significant difference between the control group and the experimental group on the degree of improvement in persuasive writing. The post hoc re-evaluation, therefore, would not have confirmed Hypothesis 5.

It is not clear why the experiment raters and the post hoc raters differed in their evaluations. There was higher inter-rater reliability among the experiment raters (pretest range .55 to .65; posttest range .39 to .55) than among the post hoc raters (pretest range .22 to .60; posttest range -.14 to .44), but neither set of raters achieved substantial agreement. A scrutiny of the ratings showed that it was not possible to identify a specific variable such as mother tongue, gender or time of rating to explain differences in evaluations. It could be argued that the post hoc ratings were more reliable because the second set of raters received the pretest and posttest scripts at the same time, which would have allowed them clearer comparisons between the two sets. On the other hand, it could be argued that the greater degree of inter-rater reliability showed by the experiment raters meant that their ratings were the more reliable. Two things, however, are clear. The first is that lack of agreement between raters casts doubt on the finding of a significantly greater overall improvement on the part of the experimental group. The second is that it seems as though practice both in ISN and PHN caused the writing to improve.

The power of the objective measures was that they did not rely on subjective evaluation. The objective changes in the texts produced by each group cannot be weakened by the awareness that the result might change if other raters evaluated the texts. Changes on all the objective measures were compared between the groups and Hypothesis 6, which stated that practice in ISN rather than in PHN would generate higher levels of accuracy in persuasive writing, was confirmed. There was a significant difference between the overall increase in the level of accuracy in persuasive writing essays between the groups where the experimental group's level of error decreased, while the control group's level increased. A further investigation to find out which types of errors had differed significantly between the groups revealed that the experimental group had made significantly fewer errors in the 'cohesion and coherence' category, specifically in the areas of reference and

punctuation. The experimental group also had significantly fewer errors in the 'other' error category, e.g. careless errors. There were no significant differences in structural changes between the groups as measured by the number and length of their t-units, error-free or otherwise, and there was no significant difference in the amount of fluency gain each group achieved.

The finding that the experimental group's level of accuracy had increased more than the control group's in persuasive writing lends support to the reliability of the experiment raters' evaluations over those of the post hoc raters. If the post hoc raters' evaluations were the more reliable than the obvious conclusion is that the increased level of accuracy shown by the experimental group over the control group did not signify improvement. This may be so, but the fact that the experimental group showed a significant decrease in the 'coherence and cohesion' and 'other' categories suggests otherwise. The findings indicate that the experimental group's text coherence had improved significantly and that they had enough processing space not to make the kind of 'careless' errors that the control group were still making.

Only further research would reveal whether the finding from this study was capable of repetition and of generalisation. The value of requiring students to exercise their imaginations and the contribution of such practice to development in writing competence is not easy to research or assess. The inclusion of one narrative writing type within another, particularly the inclusion of invention in PHN, showed that the exercise of imagination happened in any case whether it was specifically required or not. However, the extra practice in producing imagined stories that one group of students received over another over a period of months resulted in tangible improvements in that group's level of accuracy in persuasive writing.

12.4 Problems and insights

Conducting the experiment was full of pitfalls. The first difficulty arose because of the need to teach both control and experimental groups in the same room at the same time. This was not easy to do because it meant that separate group teaching was prevented, which was frustrating. It did, however,

ensure that any language teaching was given equally to both groups and helped to ensure that each group was given as nearly equal amounts of attention and encouragement as was possible. The second practical difficulty was the lack of paper for writing. It was very difficult to get the students to relax into producing untidy drafts and making mistakes when paper was so scarce. The problem was eventually resolved, but it had drastic effects on the students' willingness to try out their writing. This is the kind of problem that would presumably not occur in wealthy countries, but it is an issue that teachers need to be aware of in places where resources are scarce.

There were three difficulties to do with measurement of the writing. These were described in the previous chapter and so are mentioned only briefly here. The first was that much greater care should have been taken in designing the essay prompts, both with regard to the emotional effect they could be expected to have, as well as the effect an implicit as opposed to an explicit audience specification might have on performance. The second was that the research design had relied heavily on holistic evaluations, but despite the care taken in putting into practice advice from previous researchers, such as making sure that raters had similar levels of experience and designing a rating scale in order to standardise evaluations, it was clear that not all raters had evaluated similarly. The third difficulty was that the study had chosen to investigate only a few selected objective features of text. This raised the difficulty of establishing their relative importance with each other, as well as making it necessary to acknowledge the effect on them of text features that had not been investigated.

The study brought insights, too. The first of these was the awareness of how writing types were mixed as students developed their writing competence. Data from the student questionnaires revealed that the narrative types had been heavily mixed, although not in equal amounts. PHN had included far more invented sections than ISN had included personal histories. This may imply that as writing competence develops students gradually include increasing amounts of imagined material in their personal experience narratives. A contributing factor to the mixing of the writing types may have been the close proximity of the groups, who wanted to try out each other's essay titles, but that does not explain why the move was mainly from PHN to ISN and not in the other direction. The writing

types may have been mixed because of the culture of the writers, or because of the subjects' level of writing development, or because it always happens. The common sense conclusion is that all these factors played a part. It would seem reasonable to expect that the inclusion of one writing type within another for reasons of cognitive ease would occur more frequently at an early stage of writing development, as in the case of narrative insertions in the persuasive essays, but some of the reasons for including other writing types within the main one might always apply. Emotional need, for example, might frequently cause writers to invent in order to fulfil the requirements of an unpleasant task. Desire to impress the audience may always be present to some extent and may cause invention to provide added interest, or narrative insertions may be needed in persuasive essays in order to give examples to make a point. The awareness that writing types were mixed as writing competence developed is an important finding. More research is needed in order to monitor how much mixing usually occurs, whether it is always present to some extent, and whether it occurs less as competence develops.

The second insight was an increased awareness of how difficult it is to write. This came from my own experience in writing up the study as well as from student comments about the problems they had experienced. Students commented that writing took a great deal of time and that there was never enough time. It is easy for teachers who are teaching writing but are not engaged in writing anything themselves, to forget how difficult it is to write and to forget how long it can take sometimes just to produce a few sentences. It was clear, too, both from student comments and from an investigation of performance associated with different titles, that emotions associated with writing had powerful effects on the ease of cognitive processes and thus on performance. Negative emotional effects associated with the writing, such as feelings of guilt or fear or simply lack of success, were commented on as being demotivating. Positive emotional effects, such as the enjoyment of reading the story afterwards, or the feeling of success associated with the sense of improvement had the opposite effect. Negative emotional effects could be so powerful that the cognitive processes needed for writing ground to a halt. The drive to start or to continue writing seemed to be powered by emotion, rather than by external command, external need or will power. The writing process did,

however, bring rewards as well as difficulties. It is a common assumption that writing provokes thought and learning and this was confirmed by unsolicited comments from the students on the effects of writing on their thinking processes. They said that it made them think deeply, had provoked their imagination and that they had learned about themselves from their writing.

Another insight was that the practice of rewriting can be counter-productive. The intention behind the requirement that the students rewrite their essays after receiving comments on them was to teach them that rewriting improves essays, but in this case it did not work. The grade nine students did almost no rewriting of content when they rewrote their essays. They attempted to make surface corrections, and usually did them badly, often introducing new errors that had not been present in the first place. Their attempts at improved text organisation consisted in changing the size of paragraphs. It was possible to demonstrate to the students the logic behind the rewriting process, but they remained unconvinced so they produced the rewrites slowly, painfully and badly. It seemed that the reason students found the requirement to rewrite so onerous was that emotionally they had finished with the task. They agreed that their essays could have been improved, but had no motivation to do so. It was as if the effort of writing in the first place had tired them so much, that any further involvement with that particular piece of writing was felt to be too painful to bear. Interest in the piece had finished and motivation could not be reawakened. It seems as though the amount of rewriting that is valuable depends heavily on the level of writing maturity as well as on a writer's relationship with a particular piece of writing. For beginning writers it is demotivating and, in the case of this study, apparently counterproductive to be required to rewrite essays. Rewriting seems to work only when the writer has an ongoing emotional involvement with the text in question and feels a need to write more clearly or to change the content. Once this urge to write is finished, it is very difficult to rekindle it from the outside. When the writer is forced to rewrite without any inner need or pleasure in improving the piece, then the process appears to be counterproductive.

The most important insight of all was the renewed awareness of how important it is to listen to students. It had not been the original intention of the study to include a questionnaire or to ask

students for comments on the writing process. The questionnaires were designed as a response to the gradual awareness that writing types were being mixed. Since the aim of the experiment had been to compare practice in two kinds of narrative writing, then it was important to try and find out what had actually happened. The information gained from the questionnaires, however, is felt to contribute some of the most important findings of the study. I was impressed both by the articulate and thoughtful nature of student reflection, as well as by the fact that students clearly wished to make comments.

It became clear to me that the interpretation of 'objective' findings is not only illuminated by student perception, but that interpretations of findings on writing cannot get anywhere near the truth if student perceptions are not taken into account. Teaching and learning is a two-way affair and understanding the process must take account of both sides. In some respects the practice and development of writing is a one-way affair. It is a study of the relationship of a writer with him or herself. To deny the perception of the subject on either the writing process or the text that is produced would be to deny an essential element that contributes towards the understanding of the development of writing competence.

12.5 Implications

- It is important to listen carefully to what students have to say with the awareness that not all groups or individual students have identical experience or problems.

- Teachers should be aware that the practice of rewriting is not always helpful, especially at elementary levels of proficiency.

- The finding that fluency seemed to be more important than accuracy in the early stages of writing development suggests that teachers should give students as much as writing practice as possible in order to increase their fluency and should not worry too much about levels of accuracy in the early stages.

- There can be a wide variety of response to any title so a choice of topic should be provided in order to enable writers to perform to the best of their ability.

- Holistic evaluations should be regarded with caution since other raters' evaluations might be equally valid but different.

- Interpretations of studies which isolate specific features of texts as indicators of development or quality need to take into account the relationship between that feature and all the others, so that their contribution to the text as a whole can be assessed. Since all features of text operate together in an interaction with the reader, it is important never to lose sight of the whole piece of writing.

- The emotional effect of the essay prompt needs careful consideration since it affects motivation and therefore performance.

- Although it seems that invention may be included whatever the focus of the writing practice, the finding that those students who had been given specific practice in inventing stories had significantly increased their ability to write accurate persuasive essays suggests that it is sound pedagogic practice to require ESL students to produce imagined stories.

- The link between practice in the cognitive process of imagining and the development of writing competence was shown to be important and deserves to be further explored.

BIBLIOGRAPHY

Aitchison J. 1987. Words in the Mind. Oxford: Blackwell.

Ahai N. and N. Faraclas. 1993. 'Rights and expectations in an age of "debt crisis". Literacy and integral human development in Papua New Guinea'. In P. Freebody and A.R. Welch (eds.) Knowledge, Culture and Power. International Perspectives on Literacy as Policy and Practice. London: Falmer Press 1993: 82-101.

Ali H.A. 1989. 'A process-based approach to teaching written English to first year university students in Lebanon: an exploratory model'. PhD Thesis: University of Edinburgh.

Alderson J.C. and C. Clapham. 1992. 'Applied linguistics and language testing: a case study of the ELTS test'. Applied Linguistics 13/2: 149-167.

Alderson J.C. and Y. Lukmani. 1989. 'Cognition and reading: cognitive levels as embodied in test questions'. Reading in a Foreign Language 2: 253-270.

Allaei S.K. and U. Connor. 1991. 'Using performative assessment instruments with ESL student writers'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 227-240.

Allwright R.L., M.P. Woodley and J.M. Allwright. 1988. 'Investigating reformulation as a practical strategy for the teaching of academic writing'. Applied Linguistics 9/3: 236-256.

Arndt V. 1987. 'Six writers in search of texts: a protocol based study of L1 and L2 writing'. ELT Journal 41/4: 257-267.

Asian Development Bank. 1991. Social Indicators for Developing Countries in Asia and Australasia April 1991.

Astika G.G. 1993. 'Analytical assessments of foreign students' writing'. RELC Journal 24/1: 61-72.

Bacha N.S. and E.A.S. Hanania. 1980. 'Difficulty in learning and effectiveness in teaching transitional words: a study on Arabic-speaking university students'. TESOL Quarterly 14/2: 251-254.

Ballard B. and J. Clanchy. 1991. 'Assessment by misconception: cultural influences and intellectual traditions'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 19-35.

Barritt L., P.L. Stock and F. Clark. 1986. 'Researching practice: evaluating assessment essays'. College Composition and Communication 37/3: 315-327.

Barron C. 1986. Lexical Nativisation in Papua New Guinea English. Dept of Language and Communication Studies Research Report No 7: Papua New Guinea University of Technology, Lae.

Bartholomae D. 1980. 'The study of error'. College Composition and Communication 31: 253-269.

Bamberg B. 1983. 'What makes a text coherent?' College Composition and Communication 34/4: 417-429.

Basham C.S. and P.B. Kwachka. 1991. 'Reading the world differently: a cross cultural approach to writing assessment'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 37-49.

Benesch S. 1995. 'Genres and processes in a sociocultural context'. Journal of Second Language Writing 4/2: 191-195.

Benterrak K., S. Muecke and P. Roe 1984 Reading the Country Fremantle Arts Centre Press

Berg J. 1989. 'Metaphor, meaning and interpretation'. In A. Kasher (ed.) Cognitive Aspects of Language Use. Amsterdam: North Holland 1989: 191-205.

Bereiter C. and M. Scardamalia. 1981. 'From conversation to composition: the role of instruction in a developmental process'. In R. Glaser (ed.) Advances in Instructional Psychology Vol. 2. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bereiter C. and M. Scardamalia. 1983. 'Does learning to write have to be so difficult?' In A. Freedman, I. Pringle and J. Yalden (eds.) Learning to Write: First Language/Second Language. London:Longman 1983: 20-33.

Bereiter C. and M. Scardamalia. 1987. The Psychology of Written Composition. Hillsdale: Erlbaum

Besnier N. 1995 Literacy, emotion and authority: reading and writing on a Polynesian atoll Cambridge: CUP

Biber D. 1988 Variation across speech and writing Cambridge: CUP

Biber D. 1995 Dimensions of Register Variation Cambridge: CUP
Blanton L.L. 1994. 'Discourse, artifacts and the ozarks: understanding academic literacy'. Journal of Second Language Writing 3/1: 1-16.
Boughey C. 1997. 'Learning to write by writing to learn: a group work approach'. ELT Journal 51/2: 126-134.
Brandt D. 1989. 'The message is the massage'. Written Communication 6: 31-44.
Brewer W.B. 1988. 'Are foreign language requirements defensible in the light of recent research findings?' Language Teaching 21/4: 222.
Brice Heath S. 1983 Ways with Words Cambridge: CUP
Britton J. 1975. 'Teaching writing'. In A. Davies (ed.) Problems of Language and Learning London: Heinemann 1975: 113-133.
Britton J. 1983. 'Shaping at the point of utterance'. In A. Freedman, I. Pringle and J. Yalden (eds.) Learning to Write: First Language/Second Language. London: Longman 1983:13-19.
Brown R.G. 1991. 'Schooling and thoughtfulness'. Journal of Basic Writing 10/1:3-15.
Bruton A.S. 1981. 'A decision-making approach to the extended writing lesson'. ELT Journal 35/2: 141-146.
Byrd P. and G. Nelson. 1995. 'NNS performance on writing proficiency exams: focus on students who failed'. Journal of Second Language Writing 4/3: 273-285.
Carlisle R. and E. McKenna. 1991. 'Placement of ESL/EFL undergraduate writers in college-level writing programs'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. Ablex: New Jersey 1991: 197-211.
Carrell P. 1982. 'Cohesion is not coherence'. TESOL Quarterly 16/4: 479-488
Carrell P. 1983. 'Three components of background knowledge in reading comprehension'. Language Learning 33/2: 183-205.
Carrell P. 1985. 'Facilitating ESL reading by teaching text structure'. TESOL Quarterly 19/4: 727-752.
Carrell P. 1987. 'Content and formal schemata in ESL reading'. TESOL Quarterly 21/3: 461-481
Carrell P. and L.B. Monroe. 1993. 'Learning styles and composition'. Modern Languages Journal 77/2: 148-162.
Carroll J.B. 1983. 'Psychometric theory and language testing'. In J.W. Oller (ed.) Issues in Language Testing Research. Rowley, Mass.: Newbury House 1983: 80-107.
Carroll M. 1994. 'Journal writing as a learning and research tool in the adult classroom'. TESOL Journal 4/1: 19-22.
Carson J.G. 1992. 'Becoming biliterate: first language influences'. Journal of Second Language Writing 1/1: 37-60.
Carson J.G. and G.L. Nelson. 1994. 'Writing groups: cross-cultural issues'. Journal of Second Language Writing 3/1: 17-30.
Casanave C.R. 1994. 'Language development in students' journals'. Journal of Second Language Writing 3/3: 179-201.
Caulk N. 1994. 'Comparing teacher and student response to written work'. TESOL Quarterly 28/1: 181-188.
Chalhoub-Deville M. 1997. 'Theoretical models, assessment frameworks and test construction'. Language Testing 14/1: 3-22.
Charles M. 1990. 'Responding to problems in written English using a student's self-monitoring technique'. ELT Journal 44/4: 286-293.
Charney D. 1984. 'The validity of using holistic scoring to evaluate writing'. Research in the Teaching of English 18/1: 65-81.
Chaudron C. 1984. 'The effects of feedback on students' composition revisions'. RELC Journal 15/2: 1-15.
Chichara T., T. Sakurai and J.W. Oller. 1989. 'Background and culture as factors in EFL reading comprehension'. Language Testing 6/2: 143-151.
Chimombo H. 1986. 'Evaluating compositions with large classes'. ELT Journal 40/1: 20-27.
Choi I.C. 1992. An Application of Item Response Theory to Language Testing. New York: Peter Lang Publishing.

Christensen F. 1968. 'The problem of defining a mature style'. English Journal 57:572-579.
Clarke M.A. 1994. 'The dysfunctions of the theory/practice discourse'. TESOL Quarterly 28/1: 9-26.
Clyne M. 1981. 'Culture and discourse structure'. Journal of Pragmatics 5: 61-66.
Cohen A.D. and M.C. Cavalcanti. 1988. 'Giving and getting feedback in compositions: a comparison of teacher and student verbal reports'. Language Teaching 21/4: 244.

Cohen A.D. and M.C. Cavalcanti. 1990. 'Feedback on compositions: teacher and student verbal reports'. In B. Kroll (ed.) Second Language Writing. Cambridge: C.U.P. 1990: 155-177.
Connor U. 1991. 'Linguistic/rhetorical measures for evaluating ESL writing'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. Ablex: New Jersey 1991:215-225.
Connor U. and K. Asenavage. 1994. 'Peer response groups in ESL writing classes: how much impact on revision?' Journal of Second Language Writing 3/3: 257-276.
Connor U. and J.Linton. 1995. 'Research issues: research in the rating process. Looking behind the curtain: what do L2 composition ratings really mean?' TESOL Quarterly 29/4: 762-765.
Connors R.J. and A.Lunsford. 1988. 'Frequency of formal errors in current college writing, or Ma and Pa Kettle do research'. College Composition and Communication 39/4: 395-409.
Cooper C.R. and L. Odell. 1976. 'Considerations of sound in the composing process of published writers'. Research in the Teaching of English 10: 103-115.
Cooper C.R. and A. Matsuhashi. 1983. 'A theory of the writing process'. In M. Martlew (ed.) The Psychology of Written Language. John Wiley & Sons Ltd. 1983: 3-39.
Cortazzi M. 1994. 'Narrative analysis'. Language Teaching 27:157-170.
Crowhurst M. and G. Piche. 1979. 'Audience and mode of discourse effects on syntactic complexity in writing at two grade levels'. Research in the Teaching of English 13: 101-110.
Cumming A. 1990. 'Expertise in evaluating second language compositions'. Language Testing 7/1: 31-51.
Cumming A. 1992. 'Instructional routines in ESL composition teaching: case study of 3 teachers'. Journal of Second Language Writing 1/1: 17-35.
Cumming A. 1998. 'Theoretical perspectives on writing'. Annual Review of Applied Linguistics 18: 61-78.
Cummins J. 1983. 'Language proficiency and academic achievement. In J.W. Oller (ed.) Issues in Language Testing Research. Rowley, Mass.: Newbury House 1983: 108-129.
Damasio A.R. 1994. Descartes' Error. London: Picador.
Davidson C. and A. Tomic. 1994. 'Removing computer phobia from the writing classroom'. ELT Journal 48/3: 205-213.
Davies E.E. 1983. 'Error evaluation: the importance of viewpoint'. ELT Journal 37/4: 303-311.
Davies Samway K. 1993. '"This is hard, isn't it?" Children evaluating writing'. TESOL Quarterly 27/2: 233-258.
De Beaugrande R. 1982. 'Cognitive processes and technical writing'. Journal of Technical Writing and Communications 12/2: 121-145.
De Goes C. and M.Martlew. 1983. 'Young children's approach to literacy'. In M. Martlew (ed.) The Psychology of Written Language. John Wiley & Sons Ltd. 1983: 217-236.
Demel M. 1991. 'The relationship between overall reading comprehension and coreferential ties for second language learners of English'. TESOL Quarterly 24/2: 267-292.
Derrida J. trans. by A. Bass. 1978. Writing and Difference. London: Routledge.
Devine J., K. Railey and P. Boshoff. 1993. 'The implications of cognitive models in L1 and L2 writing'. Journal of Second Language Writing 2/3: 203-225.
Diederich P.B., J.W. French and S.T. Carlton. 1961. Factors in Judgements of Writing Ability. Princeton, New Jersey: Educational Testing Service.
Dudley-Evans T. 1994. 'Genre analysis: an approach to text analysis for ESP'. In M. Coulthard (ed.) Advances in Written Text Analysis London:Routledge 1994: 219-228.
Dyer B. 1996. 'L1 and L2 composition theories: Hillocks' environmental model and task based language teaching'. ELT Journal 50/4: 312-317.
Edge J. 1980. 'Teaching writing in large classes'. ELT Journal 34/3: 146-148.

Eisterhold-Carson J., P. Carrell, S. Silberstein, B. Kroll and P. Kuehn. 1990. 'Reading-writing relationships in first and second language'. TESOL Quarterly 24/2: 245-266.
Elbow P. 1991. 'Reflections on academic discourse: how it relates to freshmen and colleagues'. College English 53/2: 135-155.
Elsner E. 1991. 'NCTE to you'. College English 53/2:170.
Engber C.A. 1995. 'The relationship of lexical proficiency to the quality of ESL compositions'. Journal of Second Language Writing 4/2: 139-155.
Evola M., E. Mamer. and B. Lentz. 1980. 'Discrete point versus global scoring for cohesive devices'. In J. W. Oller Jr. and K. Perkins (eds.) Research in Language Testing. Newbury House 1980: 177-181.
Fathman A.K. and E. Whalley. 1990. 'Teacher response to student writing: focus on form versus content'. In B. Kroll (ed.) Second Language Writing. Cambridge: C.U.P. 1990: 178-190.
Ferris D.R. 1994. 'Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency'. TESOL Quarterly 28/2: 414-420.
Ferris D.R. 1995. 'Student reactions to teacher response in multiple-draft composition classrooms'. TESOL Quarterly 29/1: 33-53.
Fitzgerald J. and A.B. Teasley. 1988. 'Effects of instruction in narrative structure on children's writing'. Journal of Educational Psychology 78/6: 424-432.
Flahive D.S. and B.G. Snow. 1980. 'Measures of syntactic complexity in evaluating ESL compositions'. In J.W. Oller Jr. and K.Perkins (eds.) Research in Language Testing. Newbury House 1980: 171-176.
Flower L. 1979. 'A cognitive basis for problems in writing'. College English 41/1:19-37.
Flower L. and J.R. Hayes. 1981. 'A cognitive process theory of writing'. College Composition and Communication 32: 365-387.
Frankenberg-Garcia A. 1990. 'Do the similarities between L1 and L2 writing processes conceal important differences?' Edinburgh Working Papers in Applied Linguistics 1: 91-102.
Freedman A. and I. Pringle. 1980. 'Writing in the college years: some indices of growth'. College Composition & Communication 31: 311-324.
Freire P. and D. Macedo. 1987. Literacy: Reading the Word and the World. London: Routledge & Kegan Paul.
Gardner R.C. and P.D. MacIntyre. 1992. 'A student's contribution to second language learning. Part one: cognitive variables'. Language Teaching 25: 211-220.
Gardner R.C. and P.D. MacIntyre. 1993. 'A student's contribution to second language learning. Part two: affective variables'. Language Teaching 26: 1-11.
Geranpayeh A. 1994. 'Are score comparisons across language proficiency test batteries justified?: an IELTS - TOEFL comparability study'. Edinburgh Working Papers in Applied Linguistics 5: 50-65.
Golden R.M. and D.E. Rumelhart. 1993. 'A parallel distributed processing model of story comprehension and recall'. Discourse Processes 16/3: 203-237.
Goldstein L. and S. Conrad. 1990. 'Student input and negotiation of meaning in ESL writing conferences'. TESOL Quarterly 24/3: 443-460.
Gordon C.J. and C.B. Braun. 1983. 'Using story grammar as an aid to reading and writing'. Reading Teacher 37/3: 116-121.
Graves D.H. 1984. 'Patterns of child control of the writing process'. In H. Cowie (ed.) The Development of Children's Imaginative Writing. London: Croom Helm 1984: 219-232.
Green P.S. and K. Hecht. 1986. 'Reliability assessment of written communicative skills'. Language Teaching April 1986: 163-164.
Greenberg K.L. 1986. 'Review article on the development and validation of the TOEFL Writing Test: a discussion of TOEFL research reports'. TESOL Quarterly 20/3: 531-544.
Gregg V.H. 1986. Introduction to Human Memory. London: Routledge & Kegan Paul.
Grobe C. 1981. 'Syntactic maturity, mechanics and vocabulary as predictors of quality ratings'. Research in the Teaching of English 15: 75-85.
Gundlach R.A. 1982. 'Children as writers: the beginnings of learning to write'. In M. Nystrand (ed.) What Writers Know. New York: Academic Press 1982: 129-147.

Halliday M.A.K. 1985. Spoken and Written Language. Victoria: Deakin University.
Halliday M.A.K. 1994. 'The construction of knowledge and value in the grammar of scientific discourse, with reference to Charles Darwin's "The Origin of Species"'. In M. Coulthard (ed.) Advances in Written Text Analysis. 1994: 136-156.
Halliday M.A.K. and R. Hasan. 1976. Cohesion in English. London: Longman.
Hamilton J., M. Lopes, T. McNamara, and E. Sheridan. 1993. 'Rating scales and native speaker performance on a communicatively oriented EAP test'. Language Testing 10/3: 337-353.
Hamp-Lyons L. 1991a. 'The writer's knowledge and our knowledge of the writer'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 51-68.
Hamp-Lyons L. 1991b. 'Scoring procedures for ESL Contexts'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 241-276.
Hamp-Lyons L. 1991c. 'Basic concepts'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 5-15.
Hamp-Lyons L. 1991d. 'Reconstructing "Academic Proficiency"'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 127-153.
Hamp-Lyons L. 1991e. 'Pre-text: task related influences on the writer'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 87-107.
Hamp-Lyons L. 1991f. 'Issues and directions in assessing second language writing in academic contexts'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 323-329.
Hamp-Lyons L. 1995. 'Research issues: Research on the rating process. Rating non-narrative writing: The trouble with holistic scoring'. TESOL Quarterly 29/4: 759-762.
Hamp-Lyons L. and M.S. Prochnow. 1994. 'Examining expert judgements of task difficulty in essay tests'. Journal of Second Language Writing 3/1: 49-68.
Harris A. 1983. 'Language and alienation'. In B. Bain (ed.) The Sociogenesis of Language and Human Conduct. New York: Plenum Press 1983: 99-108.
Hatch E. and H. Farhady. 1982. Research Design and Statistics for Applied Linguistics. Rowley, Mass: Newbury House.
Hayes J.R. and L.S. Flower. 1980. 'Identifying the organisation of writing processes'. In L. Gregg and E.R.Steinerg (eds.) Cognitive Processes in Writing. Hillsdale, New Jersey: Lawrence Erlbaum & Associates 1980: 3-30.
Heaton J. and the Papua New Guinea Department of Education. 1985. Create and Communicate: Book 1. Papua New Guinea edition. Melbourne: Longman Cheshire.
Heaton J. and the Papua New Guinea Department of Education. 1986. Create and Communicate: Book 2. Papua New Guinea edition. Melbourne: Longman Cheshire.
Heaton J. and the Papua New Guinea Department of Education. 1987. Create and Communicate: Book 3. Papua New Guinea edition. Melbourne: Longman Cheshire.
Hedgcock J. and N. Lefkowitz. 1994. 'Feedback on feedback: assessing learner receptivity to teacher response in L2 composing'. Journal of Second Language Writing 3/2: 141-163.
Henning G. 1991. 'Issues in evaluating and maintaining an ESL writing program'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 279-291.
Hinkel E. 1994. 'Native and non-native speakers' pragmatic interpretations of English texts'. TESOL Quarterly 28/2: 353-376.
Hirose K. and M. Sasaki. 1994. 'Explanatory variables for Japanese students' expository writing in English: an exploratory study'. Journal of Second Language Writing 3/3: 203-229.
Holmes V.L. and M.R. Moulton. 1995. 'A contrarian view of dialogue: the case of a reluctant participant'. Journal of Second Language Writing 4/3: 223-251.
Homburg T.J. 1984. 'Holistic evaluation of ESL compositions: can it be validated objectively?' TESOL Quarterly 18/1:87-107.
Horning A.S. 1993. The Psycholinguistics of Readable Writing. New Jersey: Ablex Publishing Corporation.

Horowitz D. 1989. 'Function and form in essay examination prompts'. RELC Journal 20/2: 23-35
Horowitz D. 1991. 'ESL writing assessments: contradictions and resolutions'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 71-85.
Hughes A. 1989. 'Testing writing'. Testing for Language Teachers. Cambridge: C.U.P. 1989: 75-100.
Hunt K.W. 1983. 'Sentence combining and the teaching of writing'. In M. Martlew (ed.) The Psychology of Written Language. John Wiley & Sons Ltd. 1983: 99-125.
Huot B. 1990. 'Reliability, validity and holistic scoring. What we know and what we need to know'. College Composition and Communication. 41: 201-213.
Hyland K. 1990. 'Communicative competence in PNG: whose rules?' PNG TESLA Journal 7/3: 9-13.
Hyon S. 1996. 'Genre in 3 Traditions: implications for ESL'. TESOL Quarterly 30/4: 693-722.
Intaraprawat P. and M.S. Steffensen. 1995. 'The use of metadiscourse in good and poor ESL essays'. Journal of Second Language Writing 4/3: 253-272.
Ishikawa S. 1995. 'Objective measurement of low-proficiency EFL narrative writing'. Journal of Second Language Writing 4/1: 51-69.
Jacobs G. 1986. 'Quickwriting: a technique for invention in writing'. ELT Journal 40/4: 282-291.
Jacobs H., S. Zinkgraf, O. Wormuth, V. Hartfiel and J. Hughey. 1981. Testing ESL Composition: A Practical Approach. Rowley M.A.: Newbury House.
Janopoulos M. 1992. 'University faculty tolerance of NS and NNS writing errors: a comparison'. Journal of Second Language Writing 1/2: 109-121.
James C. 1974. 'Linguistic measures for error gravity'. Audio Visual Language Journal 12/1: 3-9
James C. 1977. 'Judgements of error gravities'. ELT Journal 31/2: 116-124.
John-Steiner V. and P. Tatter. 1983. 'An interactionist model of language development'. In B. Bain (ed.) The Sociogenesis of Language and Human Conduct. New York: Plenum Press 1983: 79-95
Johns A. 1986. 'Coherence and academic writing: some definitions and suggestions for teaching'. TESOL Quarterly 20/2: 247-265.
Johns A. 1990. 'L1 composition theories: implications for developing theories of L2 composition'. In B. Kroll (ed.) Second Language Writing. Cambridge: C.U.P. 1990: 24-36.
Johns A. 1991. 'Faculty assessment of ESL student literacy skills'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 167-179.
Johns A. 1995. 'Dialogue: genre and pedagogical purpose'. Journal of Second Language Writing 4/2: 181-190.
Johnson K.E. 1992. 'Cognitive strategies and second language writers: a re-evaluation of sentence combining'. Journal of Second Language Writing 1/1: 61-75.
Kamimura T. 1997. 'Composing in Japanese as a first language and English as a second language: a study of narrative writing'. RELC Journal 27/1: 47-69.
Kaplan R.B. 1966. 'Cultural thought patterns in inter-cultural education'. Language Learning 16/1 & 2: 1-20.
Kaplan R.B. 1967. 'Contrastive rhetoric and the teaching of composition'. TESOL Quarterly 1/4: 10-16.
Kaczmarek C.M. 1980. 'Scoring and rating essay tasks'. In J.W. Oller and K. Perkins (eds.) Research in Language Testing. 1980: 151-159.
Keh C.L. 1990. 'Feedback in the writing process: a model and methods for implementation'. ELT Journal 44/4: 294-304.
Khalil A. 1985. 'Communicative error evaluation: native speakers evaluation and interpretation of written errors of Arab EFL learners'. TESOL Quarterly 19/2: 335-351.
Kobayashi T. 1992. 'Native and non-native reactions to ESL compositions'. TESOL Quarterly 26/1: 81-112.
Krashen S.D. 1984. Writing: Research, Theory & Applications. Oxford: Pergamon 1984.
Kroll B. 1998. 'Assessing writing abilities'. Annual Review of Applied Linguistics 18: 219-240.

Kroll B. and J. Reid. 1994. 'Guidelines for designing writing prompts: clarifications, caveats and cautions'. Journal of Second Language Writing 3/3: 231-255.
Langer J.A. and A.N. Applebee. 1987. How Writing Shapes Thinking. Urbana, Illinois: N.C.T.E. Research Report 22.
Larsen-Freeman D. 1978. 'An ESL index of development'. TESOL Quarterly 12/4: 439-448.
Larsen-Freeman D. and V. Strom. 1977. 'The construction of a second language acquisition index of development'. Language Learning 27:123-134.
Laufer B. 1994. 'The lexical profile of second language writing: does it change over time?' RELC Journal 25/2: 21-33.
Laufer B. and P. Nation. 1995. 'Vocabulary size and use: lexical richness in L2 written production'. Applied Linguistics 16/3: 307-322.
Leki I. 1990. 'Coaching from the margins: issues in written response'. In B. Kroll (ed.) Second language Writing. Cambridge: C.U.P. 1990: 57-68.
Leki I. and J.C. Carson. 1994. 'Students' perception of EAP writing instuction and writing needs across the disciplines'. TESOL Quarterly 28/1: 81-101.
Leung C., R. Harris and B. Rampton. 1997. 'The idealised native speaker, reified ethnicities, and classroom realities'. TESOL Quarterly 31/3: 543-560.
Littlejohn A. and D. Hicks. 1989. 'Task centred writing activities'. Language Teaching 22/1: 37.
Lukmani Y. 1993a. Linguistic Accuracy versus Coherence in Assessing Examination Answers in Content Subjects. Paper presented at Language Testing Research Colloquium, Cambridge.
Lukmani Y. 1993b. Discourse Patterns and Communication Strategies in NNS Writing in Examination Answers in Economics Unpublished paper read at AILA 1993, Amsterdam.
Lumley T. 1993. 'The notion of subskills in reading comprehension tests: an EAP example'. Language Testing 10/3: 211-234.
Lumley T. and T.F. McNamara. 1995. 'Rater characteristics and rater bias: implications for training' Language Testing 12/1: 54-71.
Luria A.R. 1973. The Working Brain. Harmondsworth: Penguin Books.
Luria A.R. 1976. Cognitive Development: Its Cultural and Social Foundations. Cambridge, Mass.: Harvard University Press.
Luria A.R. 1982. Language and Cognition. John Wiley & Sons Ltd.
Luria A.R. 1983. 'The development of writing in the child'. In M. Martlew (ed.) The Psychology of Written Language. John Wiley & Sons Ltd. 1983: 237-277.
McCretton E. and N. Rider. 1993. 'Error gravity and error hierarchies'. IRAL 31/3: 177-188.
McDevitt D. 1989. 'How to cope with spaghetti writing'. ELT Journal 43/1: 19-23.
McIntyre P. 1993. 'The importance and effectiveness of moderation training on the reliability of teacher assessments of ESL writing samples'. M.A. Thesis: Melbourne University.
McKay S. 1980. 'A notional approach to writing'. ELT Journal 34/4: 308-314
McKay S. 1982. 'Literature in the ESL classroom'. TESOL Quarterly 16/4: 529-536.
Martin J.R. 1985. Factual Writing: Exploring and Challenging Social Reality. Victoria: Deakin University Press.
Martlew M. 1983. 'Problems and difficulties: cognitive and communicative aspects of writing'. In M. Martlew (ed.) The Psychology of Written Language. John Wiley & Sons Ltd. 1983: 295-333.
Mathews M. 1990. 'The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations'. ELT Journal 44/2: 117-121.
Mendonca C.O. and K.E. Johnson. 1994. 'Peer review negotiations: revision activities in ESL writing instruction'. TESOL Quarterly 28/4: 745-769.
Mlynarczyk R. 1991. 'Is there a difference between personal and academic writing?' TESOL Journal 1/1: 17-20.
Moffett J. 1968. Teaching the Universe of Discourse. Boston: Houghton Mifflin.
Morgan C. 1994. 'Creative writing in foreign language teaching'. Language Learning 10: 44-47.
Moslemi M.H. 1975. 'The grading of creative writing essays'. Research in the Teaching of English 9: 154-161.
Mullen K.A. 1980. 'Evaluating writing proficiency in ESL'. In J.W. Oller Jr. and K. Perkins (eds.) Research in Language Testing. Newbury House 1980: 160-170.

Neel J. 1988. Plato, Derrida and Writing. Carbondale & Edwardsville: Southern Illinois University Press.
Nelson G. and J. Murphy. 1993. 'Peer response groups: do L2 writers use peer comments in revising their drafts?' TESOL Quarterly 27/1: 135-141.
Nichols G. 1993. 'Selection from "The Fat Black Woman's Poems"'. In L. France (ed.) Sixty Women Poets. Newcastle: Bloodaxe Books 1993: 211-214
Nold E. and Freedman S.W. 1977. 'An analysis of readers' response to essays'. Research in the Teaching of English 11: 164-174
Nystrand M. 1982a. 'An analysis of errors in written communication'. In M. Nystrand (ed.) What Writers Know. New York: Academic Press 1982: 57-74.
Nystrand M. 1982b. 'The structure of textual space'. In M. Nystrand (ed.) What Writers Know. New York: Academic Press 1982: 75-86.
Olson D. and A. Hildyard. 1983. 'Writing and literal meaning'. In M. Martlew (ed.) The Psychology of Written Language. John Wiley & Sons Ltd. 1983: 41-65.
Oxenham J. 1980. Literacy: Writing, Reading and Social Organisation. London: Routledge & Kegan Paul.
Paltridge B. 1996. 'Genre, text type and the language learning classroom'. ELT Journal 50/3: 237-243.
Papua New Guinea Department of Education. 1995. Attainment Targets for Language, Elementary to Upper Secondary. Revised 1994, approved for trial at the 1/95 Community and Secondary Board of Studies Meeting. Unpublished Minutes. Waigani: PNG Dept of Education.
Parisi P. 1979. 'Close reading, creative writing and cognitive development'. College English 41/1: 57-67.
Pennington M.C., M.N. Brock and F. Yue. 1996. 'Explaining Hong Kong students' response to process writing: an exploration of causes and outcomes'. Journal of Second Language Writing 5: 227-252.
Pennington M.C., S. So, K. Hirose, V. Costa, J.L.W. Shing and K. Niedziefski. 1997. 'The teaching of English-as-a-second-language writing in the Asia-Pacific region: a cross-country comparison'. RELC Journal 28/1: 120-143.
Perera K. 1984. Children's Writing and Reading. Oxford: Blackwell.
Perkins K. 1980. 'Using objective methods of attained writing proficiency to discriminate among holistic evaluations'. TESOL Quarterly 14/1: 61-69.
Perkins K. 1983. 'On the use of composition scoring techniques, objective measures and objective tests to evaluate ESL writing ability'. TESOL Quarterly 17/4: 651-671.
Peterson C. 1993. 'Identifying referents and linking sentences cohesively in narration'. Discourse Processes 16: 507-524.
Phillip A. 1986. Communication Skills Needs Survey of the Papua New Guinea Provincial and National Public Service 1985 Port Moresby: Administrative College of Papua New Guinea Press.
Phillip A. 1994. 'Problems with the assessment of writing.' Paper given at the Teaching English as a Second Language Association Conference, Goroka, Papua New Guinea April 1994.
Phillip A. 1995. 'Self-alienation and the incorporation of society: the development of writing through degrees of self differentiation'. Paper presented at the University of Papua New Guinea Research Series April 1995.
Piaget J. 1972. The Principles of Genetic Epistemology London: Routledge & Kegan Paul.
Polio C.G. 1997. 'Measures of linguistic accuracy in second language writing research'. Language Learning 47/1: 101-143.
Pollitt A. and C. Hutchinson. 1987. 'Calibrating graded assessments: Rasch partial credit analysis of performance in writing'. Language Testing 4/1: 72-92.
Prucha J. 1983. 'Using language: a sociofuntional approach'. In B. Bain (ed.) The Sociogenesis of Language and Human Conduct. New York: Plenum Press 1983: 287-295.
Purves A.C. 1992. 'Reflections on research and assessment in written composition'. Research in the Teaching of English 26/1: 108-122.

Quirk R., S. Greenbaum, G. Leech, and J. Svartvik. 1972. A Grammar of Contemporary English. Harlow: Longman.
Raimes A. 1979. 'Problems and teaching strategies in ESL composition'. Language and Education: Theory and Practice 14 Center for Applied Linguistics.
Raimes A. 1987. 'Language proficiency, writing ability and composing strategies: a study of ESL college writers'. Language Learning 37/3: 439-468.
Raimes A. 1990. 'The TOEFL Test on Written English'. TESOL Quarterly 24/3: 427-442.
Raimes A. 1998. 'Teaching Writing'. Annual Review of Applied Linguistics 18: 142-167.
Regent O. 1985. 'A comparative approach to the learning of specialised written discourse'. In P. Riley (ed.) Discourse and Learning. New York: Longman 1985: 105-120.
Reid J. 1990. 'Responding to different topic types: a quantitative analysis from a contrastive rhetoric perspective'. In B. Kroll (ed.) Second Language Writing. Cambridge: C.U.P. 1990: 191-210.
Reid J. 1992. 'A computer text analysis of four cohesion devices in English discourse by native and nonnative writers'. Journal of Second Language Writing 1/2: 79-107.
Reid J. 1994. 'Responding to ESL students' texts: the myths of appropriation'. TESOL Quarterly 28/2: 273-292.
Reid J. and B. Kroll. 1995. 'Designing and assessing effective classroom writing assignments for NES and ESL students'. Journal of Second Language Writing 4/1: 14-41.
Reppen R. 1995. 'A genre-based approach to content writing instruction'. TESOL Journal 4/2: 32-35
Rifkin B. and F.D. Roberts. 1995. 'Error gravity: a critical review of research design'. Language Learning 45/3: 511-537.
Riffkin B. and F.D. Roberts. 1995. 'Error gravity: a critical review of research design'. Language Learning 45/3: 511-537.
Robb T., S. Ross, and I. Shortreed. 1986. 'Salience of feedback on error and its effect on EFL writing quality'. TESOL Quarterly 20/1: 83-95.
Rocklin E. 1991. 'Converging transformations in teaching composition, literature and drama'. College English 53/2: 177-194.
Rose M. 1984. Writers Block - The Cognitive Dimension. Carbondale & Edwardsville: Southern Illinois University Press.
Ross S., T. Robb and I. Shortreed. 1988. 'First language composition pedagogy in the second language classroom'. RELC Journal 19/1: 29-48.
Rouse J. 1979. 'The politics of composition'. College English 41/1: 1-12.
Rubin D.L. and G.L. Piche. 1979. 'Development in syntactic and strategic aspects of audience adaptation skills in written persuasive communication'. Research in the Teaching of English 13/4: 293-316.
Rumelhart D.E. 1981. 'Schemata: the building blocks of cognition'. In J.T. Guthrie (ed.) Comprehension and Teaching: Research Reviews. International Reading Assoc. Network. 1981: 3-25.
Santos T. 1988. 'Professors' reactions to the academic writing of non-native speaking students'. TESOL Quarterly 22/1: 69-90.
Santos T. 1992. 'Ideology in composition: L1 and ESL'. Journal of Second Language Writing 1/1: 1-15
Scardamalia M. and C. Bereiter. 1983. 'The development of evaluative, diagnostic and remedial capabilities in children's composing'. In M. Martlew (ed.) The Psychology of Written Language. John Wiley & Sons Ltd. 1983: 67-95.
Schoonen R., M. Vergeer and M. Eiting. 1997. 'The assessment of writing ability: expert readers versus lay readers'. Language Testing 14/2: 157-184.
Schumann J.H. 1997. 'The neurobiology of affect in language'. A Supplement to Language Learning No 48 Malden USA: Blackwell.
Scollon R. and S. Scollon 1981 Narrative, literacy and face in interethnic communication New Jersey: Ablex
Scott M.S. and Tucker R. 1974. 'Error analysis and English language strategies of Arab students'. Language Learning 24: 69-97.

Scribner S. and M. Cole. 1981. The Psychology of Literacy. Cambridge, Mass.: Harvard University Press
Sengupta S. 1998. 'Peer evaluation: "I am not the teacher'. ELT Journal 5/21: 19-28.
Shepherd V. 1994. Literature about Language. London: Routledge.
Shohamy E., C.M. Gordon and R. Kraemer. 1992. 'The effect of raters' background and training in the reliability of direct writing tests'. Modern Language Journal 76/1: 27-33.
Shrubsall P. 1997. 'Narrative, argument and literacy: a comparative study of the narrative discourse development of monolingual and bilingual 5-10-year-old learners'. Journal of Multilingual and Multicultural Development 18/5: 402-421.
Silva T. 1993. 'Toward an understanding of the distinct nature of L2 writing: the ESL research and its implications'. TESOL Quarterly 27: 657-677.
Silva T., M. Reichelt and J. Lax-Farr. 1994. 'Writing instruction for ESL graduate students: examining issues and raising questions'. ELT Journal 48/3: 197-204.
Skehan P. 1988. 'Language Testing Part 1'. Language Teaching 21/4: 211-221.
Skehan P. 1989. 'Language Testing Part 2'. Language Teaching 22/1: 1-13.
Skehan P. 1996. 'A framework for the implementation of task-based instruction'. Applied Linguistics 17/1:38-62.
Smithies M. and S. Holzknecht. 1981. 'Errors in Papua New Guinea written English at the tertiary level'. RELC Journal 12: 10-29.
Snowling M.T. 1985. Children's Written Language Difficulties. Windsor: Nelson-NREF.
Song B. and I. Caruso. 1996. 'Do English and ESL faculty differ in evaluating the essays of native English-Speaking and ESL students?' Journal of Second Language Writing 5/2: 63-182.
Spencer E., J. Lancaster, J. Rey, J. Benvie and I. McFayden. 1983. Written Work in Scottish Secondary Schools: A Descriptive Study. Edinburgh: The Scottish Council for Research in Education.
Stahl-Gemake J. and F. Guastello. 1984. 'Using story grammar with students of English as a Foreign language to compose original folk and fairytales'. Reading Teacher 38/2: 213-216.
Stansfield C. 1986. 'A history of the Test of Written English: the developmental year'. Language Testing 3/2: 224-234.
Stansfield C. and J. Ross. 1988. 'A long term research agenda for the Test of Written English'. Language Testing 5/2: 160-186.
Stevenson I. and S. Jenkins. 1994. 'Journal writing in the training of International Teaching Assistants'. Journal of Second Language Writing 3/2: 97-120.
Stubbs M. 1982. 'Language and society: some particular cases and general observations'. In M. Nystrand (ed.) What Writers Know. New York: Academic Press 1982: 31-55.
Sulaiman M. 1990. 'Newspapers: a resource for language teaching'. PNG TESLA Journal 7/3: 30-34.
Susser B. 1994. 'Process approaches in ESL/EFL writing instruction'. Journal of Second Language Writing 3/1: 31-37.
Swales J. 1990. Genre Analysis: English In Academic and Research Settings. Cambridge: C.U.P.
Sweedler-Brown C.O. 1993. 'ESL essay evaluation: the influence of sentence-level and rhetorical features'. Journal of Second Language Writing. 2/1: 3-17.
Tarone E., B. Downing, A. Cohen, S. Gillette and R. Murie. 1993. 'The writing of Southeast Asian-American students in secondary school and university'. Journal of Second Language Writing 2/2: 149-172.
Tedick D.J. 1990. 'ESL writing assessment: subject matter knowledge and its impact on performance'. English for Specific Purposes 9/2: 123-143.
Tomlinson B. 1983. 'An approach to the teaching of continuous writing in ESL classes'. ELT Journal 37/1: 7-16.
Tonkin E. 1995 Narrating our pasts: The social construction of oral history Cambridge: CUP
Upshur J.A. and C.E. Turner. 1995. 'Constructing rating scales for second language tests'. ELT Journal 49/1: 3-12.
Uzawa K. 1996. 'Second language learners' processes of L1 writing, L2 writing and translation from L1 into L2'. Journal of Second Language Writing 5/3: 271-294.

Van Bruggen J. 1946. 'Factors affecting regularity of the flow of words during written composition'. Journal of Experimental Education 15/2: 133-155.
Vann R.J., D.E. Meyer and F.O. Lorenz. 1984. 'Error gravity: a study of faculty opinion of ESL errors'. TESOL Quarterly 18/3: 427-440.
Vann R.J., D.E. Meyer and F.O. Lorenz. 1991. 'Error gravity: faculty response to errors in the written discourse of non-native speakers of English'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 181-195.
Vaughan C. 1991. 'Holistic assessment: what goes on in the rater's mind'. In L. Hamp-Lyons (ed.) Assessing Second Language Writing in Academic Contexts. New Jersey: Ablex 1991: 111-125.
Vygotsky L. 1962. Thought and Language. Cambridge, Mass: M.I.T.
Vygotsky L. 1978. Mind in Society. Cambridge, Mass.:Harvard University Press.
Vygotsky L. 1983. 'The Prehistory of Written Language'. In M. Martlew (ed.) The Psyschology of Written Language. John Wiley & Sons Ltd. 1983: 279-292.
Wagner M. 1990. 'Facing the blank page: five strategies for generating ideas for writing'. PNG TESLA Journal 7/3: 35-41.
Warshauer Freedman S. 1987. Response to Student Writing. Urbana, Illinois: N.C.T.E. Research Report No 23.
Watson C. 1982. 'The use and abuse of models in the ESL writing class'. TESOL Quarterly 16/2: 5-14.
Watson C. 1983. 'Syntactic change: writing development in the rhetorical context'. In M. Martlew (ed.) The Psychology of Written Language. John Wiley & Sons 1983: 127-139.
Whaley J.F. 1981. 'Story grammars and reading instruction'. Reading Teacher 34: 762-771.
Widdowson H.G. 1983. 'New starts and different kinds of failure'. In A. Freedman, I. Pringle and J. Yalden (eds.) Learning to Write: First Language/Second Language. London: Longman 1983: 34-37.
Wiegle S.C. 1994. 'Effects of training as raters of ESL compositions'. Language Testing 11/1: 197-223.
Wilkinson A. 1983. 'Assessing language development: The Crediton Project'. In A. Freedman, I. Pringle and J. Yalden (eds.) Learning to Write: First Language/Second Language. London: Longman 1983: 67-86.
Winfield F.E. and P. Barnes-Felfeli. 1982. 'The effects of familiar and unfamiliar cultural context on foreign language composition'. Modern Language Journal 66/4: 373-376.
Witte S. 1983. 'The reliability of mean t-unit length: some questions for research in written composition'. In A. Freedman, I. Pringle and J. Yalden (eds.) Learning to Write: First Language/Second Language. London: Longman 1983: 171-177.
Witte S.P. and L. Faigley. 1981. 'Coherence, cohesion and writing quality'. College Composition & Communication 32: 189-204.
Wolk A. 1970. 'The relative importance of the final free modifier'. Research in the Teaching of English 4: 59-68.
Wu S.M. 1995. 'Evaluating narrative essays: a discourse analysis perspective'. RELC Journal 26/1: 1-26.
Yarupawa S. 1991. 'Using Genre Analysis in determining the Discoursal Strengths and Weaknesses in the essays of first year PNG University of Technology students'. M.A. thesis: Macquarie University, Sydney.
Yarupawa S. 1994. 'Some methods used in genre analysis and their pedogogical implications'. PNG TESLA Journal 10/1: 115-124.
Yarapea A. 1990. 'Types of writing grades 5 and 6 children are capable of in PNG schools'. PNG TESLA Journal 7/3: 1-8.
Young G.M. 1985. 'The development of logic and focus in children's writing'. Journal of Language and Speech 28/2: 115-127.
Zamel V. 1982. 'Writing: the process of discovering meaning'. TESOL Quarterly 16/2: 195-209.
Zamel V. 1983a. 'Composing process of advanced ESL: six case studies'. TESOL Quarterly 17/2: 165-187.

Zamel V. 1983b. 'Teaching those missing links in writing'. ELT Journal 37/1: 22-30.
Zamel V. 1985. 'Responding to student writing'. TESOL Quarterly 19/1: 79-103.
Zhang S. 1987. 'Cognitive complexity and written production in English as a second language'. Language Learning 37/4: 469-481.
Zhang S. 1995. 'Re-examining the affective advantage of peer feedback in the ESL writing class'. Journal of Second Language Writing 4/3: 209-222.

Appendix A: Objective Predictors of Writing Development and Text Quality (Summaries of 33 Studies referred to in Chapter 2.4)

Researcher	Date	Subjects	Aim	Method	Predictors
Astika, G.G.	1993	EFL	to investigate assessment of foreign students' writing	210 samples rated by NS ESL teachers acc. to analytical scale of ESL Composit-ion Profile	*% variance in total score* 1. Vocabulary 83.75 2. Content 8.06 3. Lang. Use 4.05 4. Organisation 2.48 5. Mechanics 0.29
Carlisle & McKenna	1991	ESL/EFL/L1 undergrads	to compare ESL & non ESL trained raters & NS v NNS writing	to analyse & compare evaluations & correlate obj. measures with text quality on 3 university placement tests	• fluency (no sig diff btwn NS & NNS) • length of tu (no sig. diff btwn NS & NNS) - accuracy (total no of errors) did not influence raters significantly - there was a sig diff betwn NS & NNS on no of errors. - no diff btwn ESL trained & other raters
Casanave, C.R.	1994	ESL intermediate	to investigate lang. devel. through objective changes in journal writing	analysed writing over 3 semesters (16students) for: tu length, tu complexity, acc (error free tus)	variable • two thirds wrote longer t units • just over half wrote more accurately (but some wrote less accurately) • no increase in coordination, but fewer coordination words as beginning of sentences • content words - no steady increase, but related to topic (diversity not nec. correlated with quality)
Connor, U.	1991	ESL	to investigate measures for evaluating ESL writing	22 TWE essays were analysed acc. to several measures that were correlated with the holistic scores	6. Toulmin measure of reasoning 7. abstract vs situated style syntactic dimension 8. credibility appeals (- fluency correlation was poor, only 0.59; - correlations need to be interpreted in light of holistic score together with many other variables)

Researcher	Date	Subjects	Aim	Method	Predictors
Crowhurst & Piche (in Watson 83)	1979	L1 Gr 6 10	to see how discourse type & age affect syntactic maturity	narrative, descriptive, persuasive-written for best friend & teacher	- the more distant the audience the longer the t-units - audience has deepest impact in persuasive writing
Engber, C.A.	1995	ESL intermed - advanced	to investigate relationship of lexical proficiency to quality	66 placement essays holistically scored, then correlated with 4 lexical richness measures	1. lexical variation minus error 2. lexical variation (ratio of number of different lexical items to total number of lexical items)
Evola et al	1980	ESL uni students	to compare discrete point versus global scoring for cohesive devices	94 imagined stories analysed by subjective rating & objective score	• objective scoring of number of correct usages (- better than number of words, errors or obligatory contexts) conc - skills in use of cohesive devices are minimal indicators of lang. proficiency, & there was only weak correlation with ability to use grammatical items correctly
Ferris	1994	ESL intermed. & advanced	to see which lexical & syntactic features discriminate btwn 2 levels of ESL writing	2 levels of proficiency on 35 min placement tests analysed for fluency, vocab. syntax	*indicators of writing devel* 1. fluency 2. vocabulary 3. syntax Fluency was by far the most powerful indicator.
Flahive & Snow	1980	ESL 56 levels -collapsed to 6 levels	to see which measures of syntactic complexity & accuracy disciminate between levels of writing proficiency	3,000 ESL expository comps, analysed acc. to : length of tu, clause/tu, errors p.tu, complexity index	• length of t unit • clause/t-unit ratio errors per t-unit & complexity index did not discriminate

Researcher	Date	Subjects	Aim	Method	Predictors
Freedman & Pringle	1980	L1 college students	to find indices of growth in college writing	assignment essays (not exam) to analyse: syntax, rhet.scale, cog. measures	• level of abstracting correlates with level of education, so seems to be index of growth previously displayed rhetorical skills broke down when harder genres were attempted
Grobe	1981	L1 Gr 5 8 11	to investigate syntactic maturity, mechanics & vocab as predictors of text qual.	compared obj. measures for 3 age levels in narrative 7 expos. writing	*narrative* *expository* • length length • spelling spelling • vocab.
Homburg, T.	1984	ESL potential uni entrants	to see if holistic evals can be validated objectively	analysed levels 5/6/7 of 10 level Michigan Test for correlations of object. features	• tu length • no depend. clauses per comp. • no of EFTs per comp note uneven development: 5 6 7 fluency 193 268 265 no tus 18 21 18 connectors - unclear devel.
Hunt (in Hunt 83)	1970	L1 Gr4 8 12 skilled adults	to see if no & length of tus increase with age	count no of tus & words per tu & average for age writing type not controlled	*av no tus per sentence* Gr 4 - 1.6 decrease 8 - 1.4 with age 12 - 1.2 bec. better adults - 1.2 punctuation. *av no words per tu* Gr 4 - 8.6 increase 8 - 11.4 with age 12 - 14.4 bec. noun adults- 20.3 modifiers & nominalistions increase.
Intaraprawat & Seteffensen	1995	ESL uni students	to investigate use of meta-discourse features in good & poor essays	analyse good & poor persuasive essays for meta-discourse features, fluency, no of tus.	• density of meta-discourse features discriminated well between good and poor essays • fluency • no of t-us

Researcher	Date	Subjects	Aim	Method	Predictors
Ishikawa, S.	1995	EFL low proficiency	to compare benefit of: answering qs on picture story or holistic writing out of picture story	stories analysed acc. to 24 measures to see which discriminated between low levels of proficiency	1. length of error-free clauses 2. no of error-free clauses per composition discriminated between low levels of proficiency wiith small differences
Kamimura, T	1997	EFL	to compare L1-Japanese and L2 compositions to see which obj. features are related to quality	narrative compositions analysed acc. to no of sentences/ words/idea units & correlated with holistic scores	1. fluency 2. no of idea units Japanese & English writing correlated once certain threshold level passed.
Larsen-Freeman	1978	ESL 5 levels of uni students	to develop ESL index of writing growth	analysed 212 comps for fluency, length of tu, % EFTs, EFT length	*best disciminators over 5 proficiency levels* • % of error free t-units • av. length of error free tu
Laufer, B.	1994	ESL advanced	to see if lexical profile of ESL writing changes over time	plot progress of lexical profile & lexical variation at 3 points over 1 academic yr	• lexical richness increased a little (Lexical Frequency Profile -how many most used, less used and least used words) no progress in lexical variation within essays, & no correlation btwn the 2 measures
Miller (in Watson 83)	1980	L1 freshman	to see how level of fluency changes with discourse type	descriptive, explanatory, persuasive writing compared on fluency measure	level of fluency in descriptive & explanatory writing lost when asked to write persuasive essay on abstract topic to distant audience (fluency returned when subjects given persuasive writing on concrete topic to close audience)

Researcher	Date	Subjects	Aim	Method	Predictors
Mullen, K.A.	1980	ESL	to see which text features predicted overall quality	117 essays rated by 5 pairs of raters on holistic scales for: structure, organisation quantity, vocabulary	• best - vocabulary poorest - organisation (sig. difference btwn one pair of raters & differences between others
Nold & Freedman	1977	L1 freshman	to analyse readers' response to essays	22 subjects wrote 4 pw essays over several months - anal. to see which obj. features indicate devel. & quality	• modifiers, esp. final free modifiers • fluency • vocabulary -length of tu not a predictor -topic affected amount written, but not significantly
Perkins	1980	L2 advanced	to find which obj. measures contribute to text quality	to correlate objective measures with hol. evaluations	• EFTs per comp. • length of EFTs • errors per tu • total errors (not fluency or tu measures without accuracy)
Polio	1997	ESL undergrads postgrads	to see which linguistic measures of accuracy could be used reliably	38 essays rated holistically, counted EFTs & no & type of error	*reliable* EFTs error counts *unreliable* holistic scale
Rubin & Piche	1979	L1 Gr4 8 12 skilled adults	to see if audience differences affect syntax & strategies	persuasive writing to 1) intimate other, 2) less known other 3) reader of newspaper	• syntactic complexity increased consistently with age • low intimacy - longer clauses • intermediate intimacy - great variety of appeals conc - diffs due to audience can be as great as diffs due to age

Researcher	Date	Subjects	Aim	Method	Predictors
Scott & Tucker	1974	ESL low intermediate	to investigate errors for cause & relationship to writing develop-ment	22 subjects wrote a few sentences about 3 pictures at 2 points during term	• fewer errors with: finite verbs/ preps/ repetition of subj & obj./ rel.clauses/ pronouns - SVA & errors with articles did not change much. - Errors caused by IL interference not L1.
Sweedler-Brown, C.O.	1993	L1 & ESL intermediate	to compare influence of sentence-level and rhetorical features on ratings	ESL essays corrected for sentence level errors. then original ESL, corrected ESL, & L1 essays rated	• accuracy (corrected ESL essays rated higher) • holistic analytical scores correlated with overall eval on sentence level features, & grammar/mechanics no correlation on analytic scores for rhetorical org/para. devel.
Tarone et al	1993	ESL Gr 8 S.E.Asia 10 12 uni students L1 uni students	to compare features of ESL writing at various grade levels & features of L1 uni writing	used same pers. hist. narrative topic for all subjects to compare analytical scale scores on features of writing	*E.S.L. grade level* accuracy fluency no sig. difference organis. on any measure coherence *yrs in U.S & age on arrival* accuracy fluency sig. difference organis. on all measures coherence *ESL v. L1* accuracy fluency sig. difference organis. on all measures coherence
Vann et al	1984	ESL	to find hierachy of errors & what factors influence response	164 Iowa Uni profs rate 12 typical ESL errors in 24 sentences	*hierarchy of error* 1. it deletion/tense/word order/rel clause errors 2. preps/pron.agrmnt/sva 3. spelling/articles *age affects response* • most tolerant - 34 & under/55+ • least tolerant - 45-54 *discipline affects response* so.scientists more tolerant

Researcher	Date	Subjects	Aim	Method	Predictors
Vann et al	1991	ESL	to find hierarchy of errors in cont. discourse & what factors influence response	analysed response by uni faculty to 3 types of error in (doctored) essays: articles spelling verb form	*hierarchy of errors* 1. verb forms 2. article errors 3. spelling *factors affecting response* gender - women less strict discipline -so.sci. less strict age - not statist. sig. -response to errors complex not solely controlled by quality or quantity or error
Watson (in Watson 83)	1979	L1 high school adv.college Ss	to see how syntax changes with age & discourse type	expressive, persuasive-acc. to 17 syntactic features	best measures: 1)free modifiers -global 2)final free modifiers-global 3) mean t-unit length *expressive* high sch - 12. 05 college - 14.18 *persuasive* no difference 4) non-clause adjective modifiers
Witte & Faigley	1981	L1 freshman	to see which features of text indicate quality	top 5 & bottom 5 of 90 essays for sentence combining analysed acc. to: error, syntax, no &type of cohesive tie	• overall accuracy (voc. related to ability to invent) • fluency • mean t-unit length & clauses • non-restrictive modifiers cohesive ties (profiles show important differences btwn invention skills of poor & good writers)
Wolk, A.	1970	L1coll.students prof. writers	to see how syntax changes with skill level	analysed writing for t-units & free modifiers	*av. t-unit length* student - 15.5 professional - 16.6 *free modifiers(% of total words)* stud. prof. initial 8.7 9.7 medial 3.8 7.6 final 8.4 13.7 total 20.9 31.0 - % of final free modifiers discriminates most

Researcher	Date	Subjects	Aim	Method	Predictors
Zhang, S.	1987	ESL intermed.	to see how cognitive complexity of q. affects written response	63responses to 2 levels of cognitive complexity q. rated for: fluency, syntax, accuracy.	*discriminators btwn 2 levels* • fluency • syn. complexity (sentence length & clauses per sentence) (note: not accuracy)

Appendix B: Pretest Prompts

PHN

1. A HOUSE BUILDING CELEBRATION I WILL ALWAYS REMEMBER

Describe a house building celebration in your place that you particularly remember. What happened? Why was it memorable?

2. A HARVEST CELEBRATION I WILL ALWAYS REMEMBER

Describe a harvest celebration in your place that you particularly remember. What happened? Why was it memorable?

3. A BRIDE-PRICE CELEBRATION I WILL ALWAYS REMEMBER

Describe a bride-price celebration in your place that you particularly remember. What happened? Why was it memorable?

ISN

1. A DAY IN THE LIFE OF A BIRD

Imagine that you are a bird. Say what you look like and where you live and describe what happened to you yesterday.

2. A DAY IN THE LIFE OF A FISH

Imagine that you are a fish. Say what you look like and where you live and describe what happened to you yesterday.

3. A DAY IN THE LIFE OF A PIG

Imagine that you are a pig. Say what you look like and where you live and describe what happened to you yesterday.

PW

1. either VIOLENT FILMS SHOULD NOT BE SHOWN ON TV.
   or VIOLENT FILMS SHOULD BE SHOWN ON TV.

Decide what you think and choose ONE of these titles. Write about why you think violent films should or should not be shown on television.

2. either PEOPLE SHOULD BE FORCED TO PAY A FINE FOR THROWING RUBBISH ON THE STREETS
   or PEOPLE SHOULD NOT BE FORCED TO PAY A FINE FOR THROWING RUBBISH ON THE STREETS

Decide what you think and choose ONE of these titles. Write about why you think people should or should not be forced to pay a fine for throwing rubbish on the streets.

3. either ALCOHOL SHOULD BE BANNED IN PNG
   or ALCOHOL SHOULD NOT BE BANNED IN PNG

Decide what you think and choose ONE of the titles. Write about why you think alcohol should or should not be banned in PNG.

Appendix C: Posttest Prompts

PHN

1. MY FIRST SCHOOL FRIEND

Tell the story of how you met your first school friend. Write about what you did together the first day you got to know each other.

2. THE BEST PRESENT I EVER RECEIVED

Tell the story of the best present you ever received. Describe the present and what you did with it. Explain why it meant so much to you.

3. MY PUNISHMENT

What was the worst punishment you were ever given? Why were you given this punishment? What happened? How did you feel afterwards?

ISN

1. MY SECRET FRIEND

Tell the story of how you met a dog that talked. The dog talked only to you, not to other people. Describe how you first met this animal and how the dog became your secret friend.

2. AN UNUSUAL PRESENT

One day a small parcel arrived in the mail for you. Once the outer wrapping was removed, you found a small cardboard box. Inside the box you found a shiny silver ball with 9 tiny knobs on top of it. Explain what this unusual present turned out to be, and what you did with it.

3. THE ROYAL PUNISHMENT

You are Queen (or King) of a large country and your advisor, whom you trusted, has tricked you. Tell the story of how you punished him.

PW

1. SETTLEMENT IN URBAN AREAS

Write to the *Post Courier* explaining either why you agree or why you disagree with the following statement: 'People should not be allowed to settle in the urban areas if they have no job or other means of support in that place.'

2. THE RIGHT TO CHOOSE

Write to the *Post Courier* explaining either why you agree or why you disagree with the following statement: 'Young people should have the right to make their own choice of marriage partner without interference from their parents.'

3. PENALTIES FOR BREAKING ROAD SAFETY LAWS

Write to the *Post Courier* explaining either why you agree or why you disagree with the following statement: 'There should be severe penalties for anyone who breaks the new road safety laws, such as driving under the influence of alcohol, for example, or refusing to make their passengers wear seat belts.'

Appendix D: Holistic Impression Rating Scale

All essays are to be rated on a scale of 0 - 5.

5 - excellent
4 - good
3 - average
2 - below average
1 - poor
0 - very poor

5 - excellent

Organisation & Clarity - The essay is well-organised and there is a sense of development that is easy to follow from beginning to end. The meaning is clear even though the story or argument may be detailed.
Interest - The essay is interesting because it contains plenty of detail (narrative) or logical arguments with explanations (persuasive writing). It is enjoyable and memorable.
Accuracy - The essay is accurately written with only one or two minor errors of grammar or spelling. Punctuation is used correctly.

4 - good

Organisation & Clarity - The essay is well-organised and easy to understand. There is a clear sense of development from beginning to end.
Interest - The essay is interesting because it contains story detail (narrative) or good arguments with explanations (persuasive writing).
Accuracy - The essay is generally accurate with only minor errors of grammar and spelling. Punctutation is used correctly.

3 - average

Organisation & Clarity - The essay is generally clear but there is a lack of development.
Interest - The essay is only moderately interesting because the story (narrative) or arguments used (persuasive writing) tend to be very ordinary.
Accuracy - The essay is fairly accurate although there are some errors of grammar and spelling. Most of the punctuation is used correctly.

2 - below average

Organisation & Clarity - The essay is not well organised. It contains parts that are vague and confused, or the essay is simplistic. There is no clear development throughout the essay.
Interest - The essay is either short and simplistic or it is vague and confusing, so the essay is not interesting.
Accuracy - The essay contains quite a few errors of grammar or spelling. The punctuation is sometimes lacking.

1 - poor

Organisation & Clarity - The essay is either confused and very difficult to follow, or it is oversimple and extremely short.
Interest - The essay is too short or too confused to be interesting.
Accuracy - The essay may be fairly accurately written if it is very short. Otherwise there are many errors. There is a lack of punctuation.

0 - very poor

Organisation & Clarity - The essay is either impossible to follow, or consists of one sentence or less, usually the latter.
Interest - There is not enough written to be interesting.
Accuracy - There is very little language to assess, or the language contains errors of all kinds.

Appendix E: Error Categories

Category	Error Type		Example
Vocabulary	verb	- wrong one	The things which the man usually *spent* were things like flour, rice etc.
	noun	- wrong one	I am a pig and people don't like *heifers*.....
	adjective	- wrong one	...kaukau [sweet potatoes] were *obese*.....
Grammar	noun (wrong form)	- plural problems	... all the *peoples* on the beach...
			Three *woman* came....
		- un/countable	...I recived many *informations*...
		- other	... to seek *assistant*...
	verb (wrong form)	- subject/verb agreement	Look, he *walk* like dancer....
			.
		- tense & aspect	..they *work* and sang all night long.....
			..he had *being* there long time.
		- voice	..it *cook* now, its ready.....
		- other (e.g. infinitives, auxiliaries etc.)	...I told him *go* down there...
	pronoun (wrong form)	- reflexive	...I just wanted to protect *me*.....
	adjective (wrong form)	- possessive	.....*he's* head was high....
		- demonstrative	..*this* people are all same.....
		-comparative	...it was *more easy* than I thought.
		- other	To be *simplicity*, I think.....
	adverb (wrong form)		..she did it *easy*....
			.
	articles	- wrong	..at *a* same time....
		- omitted	...soo*n* *f*ood was ready....
		- unnecessary	... ...he was giving *a* very good support....
		- a/an	....bring it to *a* end.....
	redundancy		They wore traditional costumes that I saw *them*..
			The lady's family usually put money to spend *it* on the man.
	prepositions	- wrong	You are responsible *of* this..
		- omitted	...and it contributed the occasion..
	word order		

Cohesion & Coherence	reference	-pronoun	When they have a special celebration, which means people getting *them* for and celebrating *them* with the other clan people..
		- clause	....rascals did murders, which they bashed up wives....
		- missing or logically confused	Clans also cooperated to make *the same thing* which the whole village made.
	omission	- main verb	One very bad thing about showing violent films [is that] it encourages people to do them.
		- any item necessary for sense	
	logic		It also gives poor to the people who are rich.
	punctuation	- full stop/capital letter - question mark - comma - apostrophe - other	
	conjunctions		We used to check our kaukau to see *that* they were growing well. They just throw the piece of food they were eating on the ground *and while* the rubbish bins are a meter or two away from the.
Spelling			- not one *amoung* them....
Other	style carelessness (or a mistake in a word that had previously been used correctly)		I *encountered* her in the store.. I was *no* sure... *The* were happy....
	one word or two		..there were *alot* of leaves...

Appendix F: Named Questionnaire and List of Treatment Titles for each Group

NAMED QUESTIONNAIRE - GROUP 1 (PHN)

First Name(s)_____
Surname_____
Class_____ Date_____

Part 1

Check each essay in your exercise books and decide whether it was
T true - a retelling of your own experience
PT partially true - a retelling of your own experience with some imagined bits
SS somebody else's story - a story that you had heard or read or seen on television
I imagined - an imagined experience (it did not actually happen)

Write down any comments you have about that particular essay - e.g. you liked it, didn't like it, wrote it while you felt sick etc. If you do not have any comments, that is fine.

1. Escape from Danger _____
Comments:_____
_____
_____

2. My Life Story
3. The Worst Thing I ever Did
4. The First Time I Watched Television
5. An Exciting Journey
6. Child Minding
7. A Mysterious Place
8. A Day in the Life of a Provincial High School Student
9. A Funny Thing
10. A Friend in Need
11. The Best Letter I Ever Received
12. The Storm
13. The Bad Deed
14. My Revenge
15. An Exciting Ride
16. A Fishy Story
17. My Handicapped Friend
18. A Frightening Experience
19. Hurt in an Accident
20. A Memorable Shopping Trip

Part 2

Answer the following questions as honestly as you can.
1. Which was your favourite essay?
2. Why?
3. Which essay did you like least?
4. Why?
5. Would you rather have been in Group 2?
6. Why or why not?

NAMED QUESTIONNAIRE - GROUP 2 (ISN)

First Name(s)_____
Surname_____
Class_____ Date_____

Part 1

Check each essay in your exercise books and decide whether it was
I totally imagined - an imagined story invented by yourself
PI partially imagined - a mixture of real and imagined experience
SS somebody else's story - a story that you had heard or read or seen on television
T true - your own experience

Write down any comments you have about that particular essay - e.g. you liked it, didn't like it, wrote it while you felt sick etc. If you do not have any comments, that is fine.

1. Escape from the Sea _____
Comments:_____
_____
_____

2. The Life Story of the Most Beautiful Person in the World
3. The Day I Robbed the Bank
4. My First Television Appearance
5. A Trip to Midwinkle
6. Looking after Colin
7. The House of Happiness
8. A Day in the Life of a Prime Minister
9. The Teacher who Made Us all Laugh
10. The Letter that Changed my Life
11. Ada, the Helpful Spirit
12. The Storm that Destroyed Papua New Guinea
13. The Wicked Woman
14. Mr Tapoi's Revenge
15. My First Driving Lesson
16. The Mermaid
17. Blind!
18. The Night Bird
19. Buried Alive!
20. One Million Kina Shopping Spree

Part 2

Answer the following questions as honestly as you can.

1.. Which was your favourite essay?
2. Why?
3. Which essay did you like least?
4. Why?
5. Would you rather have been in Group 1?
6. Why or why not?

THE WORST THING I EVER DID

The stealings of cucumbers was the worst thing I ever did. It took place in the school garden which is near the river. I believed that no-thing was going to happen to us. But why did it and how did the school find out that the other two boys and I were the ones who stole the cucumbers? Well, we have to find out why and how we were found guilty of stealing the cucumbers.

It was on Tuesday, the seventh day of May, last year in the afternoon when myself and the other two boys decided to go in the garden of eden of Laloki and there we could see a lot of precious cucumbers with our eyes wide open. They were ready to be harvested and we were asked not to touch anything like cucumber in the garden but we have broken the law because nothing like cucumber can not stop us from getting it.

Meanwhile, I started off by taking out two cucumbers from the garden and that made my two boys confident to accompany me to also harvest the cucumbers. I was confident that nothing would happen to me and the two boys. We took as many cucumber as we could in the garden. While getting the cucumbers, I told the two boys to keeps their eyes open on to the road to see if anybody was in sight. And they did what I asked them to do.

However, some time later, I felt something was going to happen to me and my two colleagues so I told them to hurry up with what they were doing. When I looked back, I saw the headmaster on the road and my heart started to beat up slowly and faster. I quickly ran in the bush just nearby the garden and hid myself in there. Unfortunately, it was too late, the headmaster had seen us already so he called us up and told us to see him in his office.

After he had seen us in his office, he asked us to stand up in front of the assembly ground for the all students to have a good look at us! However, after being seen, we were put on punishment by eating the all leaves of cucumber which we had taken from the garden. I was really knocked out when I ate those leaves. We ate and ate for two hours in the office. Meanwhile, life at that particular time was hard and angry at the same time.

While we were in the office, I recalled what I was doing in the first place and really regrett

Appendix H: Anonymous Questionnaire

ANONYMOUS QUESTIONNAIRE

1. Did you find the essay corrections helpful?
2. Did you find the comments at the end of the essay helpful?
3. Do you think your essay writing has improved?
4. Did you enjoy the writing project?
5. What was the best thing about the writing project?
6. What was the worst thing about the writing project?
7. Any further comments:

Appendix I: Differences between Writing Types on Objective Measures

ANOVA					
	PHN	ISN	PW	F	p<0.05
n	(68)	(68)	(68)		
	mean	mean	mean		
Grammatical Structure					
t-units per 100 words	7.36	8.58	6.17	36.74	0.000*
words per t-unit	14.87	12.00	17.09	34.55	0.000*
error-free t-units per 100 words	3.15	4.40	1.54	57.12	0.000*
words per error-free t-unit	11.63	10.25	12.87	9.82	0.000*
Fluency (av. words per essay)	275.84	268.60	232.72	4.62	0.011*
Accuracy (errors per 100 words)					
Vocabulary	0.324	0.524	0.909	11.90	0.000*
Grammar					
wrong form	2.676	2.288	2.616	1.03	0.36
articles	0.607	0.557	0.649	0.25	0.776
redundancy	0.263	0.384	0.378	1.53	0.219
prepositions	0.237	0.303	0.435	3.59	0.029*
Total	3.784	3.518	4.078	1.05	0.352
Cohesion & Coherence					
reference	0.275	0.075	0.677	23.26	0.000*
omission	0.647	0.699	0.968	4.02	0.019*
punctuation	1.612	1.422	1.679	0.53	0.591
conjunctions & logic	0.146	0.227	0.207	0.98	0.376
Total	2.680	2.422	3.531	6.12	0.003*
Spelling	0.897	0.879	1.153	1.89	0.154
Other	0.374	0.350	0.344	0.09	0.911
TOTAL	8.057	7.709	10.015	5.91	0.003*
*significant difference (p<0.05)					

TITLE – either ALCOHOL SHOULD BE BANNED IN PNG
or ALCOHOL SHOULD NOT BE BANNED IN PNG
Decide what you think and choose ONE of the titles. Write about why you think alcohol should or should not be banned in PNG.

Write as much as you can in the time available.

Alchohol Should be banned in PNG
because PNG is a developing nation
and alchohol is also a drug and
will cause many problems to the
nation and the people. By this
I mean People getting drunk and
throwing bottles every where (littering)
while they are under liquor
which makes the town, village
etc filthy (broken bottles every where)
and its a fact that people
who are under liquor on a moving
vehicle etc always fall out
accidents. And if they are alive
its a warning but if died then its
another problem arising in the
nation by means of curbing
the growth of the population.
And the main problem is

now adays most of the young children (esp boys) and some girls maybe are discouraged by this alchohol. They are still in school when they keep in ~~feel~~ touch with this alchohol their spirit of schooling and their abolities change. And its the waste of parents money.

And also achohol can cause problems amongn the families. Man being drunk, and belted up the wife and from their separation or other problems arise

A small thing like a liquid (ALCHOHOL) but it can play up with people and lead them to problems etc.

So I think alchohol should be banned in PNG for the benefit of young children in schools families and for the good of the nation as a whole.

TITLE - MY PUNISHMENT

What was the worst punishment you were ever given? Why were you given this punishment? What happened? How did you feel afterwards?

Write as much as you can in the time available.

The worst punishment I had was when I was in Grade Seven. It was not only me who did the punishment but there were some other students too. One day one of our subject teachers came to our classroom and gave us some work to do.

Soon his time was up so he left the classroom to go to the next class. But before he left he gave us some more work together with the work we did during the day.

But there was a lot to do because he was not the only teacher who left us with home-work. so we I found out there wasn't enough time for me and the other students to complete our home--work.

The next day he came in expecting answers from the students for the homework he gave us. He started teaching and then went on to the questions for homework expecting the answers but the response was very poor. He asked all the students if

they had done the homework but most said no. He got very angry when he heard this.

Then he thought of what punishment to give to us. And then he left the classroom returning with a box of papers and distributed 20 to each student. And after this he told us to write "I must do my homework" until all the papers were full.

I wrote and wrote but couldn't finish and then I was sent to the school piggery to clean it out. I was given some scrubbing brushes and some other things to clean the piggery out.

I started working but I was alone so I dropped tears while working. I washed the floor and removed the pigs' waste and then washed the pigs. After washing the pigs I went on to wash the floor again.

After cleaning the piggery I went to the teacher told him about my work. And then went to the dormitory to rest. But after all that hard work I did I said to myself I must complete the homework given to me by my teachers and never do such punishment again.

And from then on I also kept away from punishment and made that my last punishment

TITLE - MY SECRET FRIEND
Tell the story of how you met a dog that talked. The dog talked only to you, not to other people. Describe how you first met this animal and how the dog became your secret friend.

Write as much as you can in the time available.

One fine day I decided to go for a walk on the beach, before I lelf I took some fruits food and water with me, On my way To the beach. I took everything that could help me on the beach, and there was no body to help me carry my things down. I sat down and thought for a while, but I couldn't come up with a succeessful idea, afterwards a thought came into mind, it was to carry my belongings at a time.

Any where I did not notice something which was staring at me with a helpful and happy face, so and that was my dog. I was so happy that I send hello to the dog and to my surprise the dog answered me, I was very very surprise and I couldn't believe it, some minutes later the dog asked me if I was in need of help.

I told the dog my problem and he told me not to worry, the dog helped me carry my things down to the beach, the dog talked and laughed and I was so happy with my dog.
When people came closer the dog wouldn't talked to them he was afraid. We played and washed together when ever people came closer the ~~th~~ dog kept silent. And we also ate, slept cracked jokes together.

After our day at the beach we returned home in the afternoon. The was always with me where ever I went, we ~~always~~ talked played together and I really loved my dog because my dog was the only ~~animal~~ that could really to a human being like me, and I'm so proud of my dog.

I was very happy from that day ~~on~~ wards. the day I met and heard the dog talking I so happy to find ~~a~~ a such dog that could talk like a human being!! What a lucky to have a such dog like ~~#~~ spunky. The dogs name was spunky.

Appendix K: Change Over Time on Objective Measures

1) PHN

n=68
t-tests

	pretests	posttests	change	t	p
Grammatical Structure					
t-units (per 100 words)	7.36	8.51	+1.15	3.68	0.0002*
words per t-unit	14.87	12.14	-2.73	-4.90	0.000*
error-free t-units (per 100 words)	3.14	4.51	+1.37	4.94	0.0000*
words per error-free t-unit	11.46	10.40	-1.06	-2.15	0.033*
Fluency (av. words per essay)	275.8	377.5	+101.7	6.31	0.0000*
Accuracy (average errors per 100 words)					
Vocabulary	0.32	0.54	+0.22	-2.66	0.0087*
Grammar:					
wrong form	2.68	2.11	-0.57	1.91	0.058
article	0.61	0.35	-0.26	2.73	0.0075*
redundancy	0.26	0.14	-0.12	2.59	0.011*
prepositions	0.24	0.23	-0.01	0.06	0.95
Total	3.78	2.83	-0.95	2.67	0.0086*
Cohesion & Coherence:					
reference	0.28	0.09	-0.19	3.47	0.0008*
omission	0.65	0.75	-0.1	-0.97	0.34
punctuation	1.61	0.67	-0.94	4.87	0.0000*
conjunctions & logic	0.15	0.14	-0.01	0.25	0.80
Total	2.68	1.64	-1.04	3.99	0.0001*
Spelling	0.90	0.79	-0.11	0.76	0.45
Other	0.37	0.42	+0.05	-0.68	0.50
TOTAL	8.06	6.23	-1.83	3.04	0.0029*

*significant ($p<0.05$)

2) ISN

n=68 t-tests	pretests	posttests	change	t	p
Grammatical Structure					
t-units (per 100 words)	8.58	8.19	-0.389	1.41	0.16
words per t-unit	12.00	13.08	+1.08	1.96	0.053
error-free t-units (per 100 words)	4.40	4.22	-0.18	-0.60	0.55
words per error-free t-unit	10.25	10.64	+0.39	0.91	0.36
Fluency (av. words per essay)	268.6	368.3	+99.7	6.72	0.0000*
Accuracy (errors per 100 words)					
Vocabulary	0.52	0.43	-0.09	1.05	0.30
Grammar:					
wrong form	2.29	2.14	-0.15	0.58	0.56
article	0.56	0.41	-0.15	1.48	0.14
redundancy	0.38	0.21	-0.17	2.42	0.017*
prepositions	0.30	0.22	-0.08	1.37	0.17
Total	3.53	3.00	-0.53	1.50	0.14
Cohesion & Coherence:					
reference	0.08	0.11	+0.03	-1.02	0.31
omission	0.70	0.72	+0.02	-0.18	0.86
punctuation	1.42	0.73	-0.69	3.48	0.0008*
conjunctions & logic	0.23	0.16	-0.07	1.42	0.16
Total	2.76	1.8	-0.96	2.83	0.0056*
Spelling	0.88	0.98	+0.1	-0.65	0.52
Other	0.35	0.39	+0.04	-0.59	0.56
TOTAL	8.58	7.01	-1.57	2.25	0.026*

* significant ($p<0.05$)

3) PW

n=68 t-tests	pretests	posttests	change	t	p
Grammatical Structure					
t-units (per 100 words)	6.18	5.79	-0.39	-1.78	0.078
words per t-unit	17.09	18.02	+0.93	1.40	0.16
error-free t-units (per 100 words)	1.54	1.77	+0.23	1.16	0.25
words per error-free t-unit	12.87	13.59	+0.72	1.03	0.31
Fluency (av. words per essay)	232.7	291.0	+58.3	4.36	0.0000*
Accuracy (errors per 100 words)					
Vocabulary	0.91	1.28	+0.37	-2.15	0.033*
Grammar:					
wrong form	2.62	2.78	+0.16	-0.63	0.53
article	0.65	0.57	-0.08	0.66	0.51
redundancy	0.38	0.26	-0.12	1.60	0.11
prepositions	0.44	0.56	+0.12	-1.39	0.17
Total	4.08	4.17	+0.09	-0.28	0.78
Cohesion & Coherence:					
reference	0.68	0.41	-0.27	2.22	0.029*
omission	0.97	0.81	-0.16	1.02	0.31
punctuation	1.68	1.12	-0.56	2.55	0.012*
conjunctions & logic	0.21	0.16	-0.05	0.78	0.44
Total	3.53	2.5	-1.03	2.95	0.0037*
Spelling	1.15	1.6	+0.45	-2.46	0.015*
Other	0.34	0.79	+0.45	-4.87	0.0000*
TOTAL	10.01	10.33	+0.32	-0.42	0.67

* significant ($p<0.05$)

Appendix L: Practice Effect of ISN on performance in PHN and ISN

1) Comparison of the groups on holistic scores

PHN

Gain scores of the control group and the experimental group were compared with an unmatched t-test.

	Control Group (PHN)	Experimental Group (ISN)		
n	(34)	(34)	t	p<0.05
	mean	mean		
	+1.91	+2.03	-0.25	0.81
PHN gain scores compared by an unmatched t-test				

There was no significant difference between the groups on improved performance in PHN.

I SN

Gain scores of the control group and the experimental group were compared with an unmatched t-test.

	Control Group (PHN)	Experimental Group (ISN)		
n	(34)	(34)	t	p<0.05
	mean	mean		
	+1.26	+2.26	-2.00	0.049*
ISN gain scores compared by an unmatched t-test				

The experimental group improved significantly more than the control group in ISN.

2. Comparison of the groups on objective measures

2a Grammatical structure

PHN

	Control Group	Experimental Group	t	p<0.05
n	(34)	(34)		
	mean	mean		
efts (per 100 words)	1.05	1.57	-1.22	0.23
words per eft	-0.75	-1.22	0.53	0.60
t-units (per 100 words)	0.91	1.29	-0.77	0.45
words per t-unit	-2.45	-2.54	0.08	0.94
unmatched t-test to compare PHN gain scores				

There were no significant differences between the groups in change of PHN structure.

ISN

	Control Group	Experimental Group	t	p<0.05
n	(34)	(34)		
	mean	mean		
efts (per 100 words)	+0.33	-0.68	2.06	0.044*
words per eft	-0.05	+0.73	-1.11	0.27
t-units (per 100 words)	+0.06	-0.83	1.89	0.064
words per t-unit	+0.20	+1.95	-1.95	0.055
unmatched t-test to compare ISN gain scores				

There was a significant difference between the groups in the number of error-free t-units (efts). The experimental group's development resulted in an average decrease of just over two thirds of an error-free t-unit, while the control group's development resulted in an average increase of a third of an

error-free t-unit. The experimental group showed an increase in number of words per error-free t-unit over the control group, but the difference was not significant.

In the traditional measures, there was a difference approaching significance in number of words per t-unit, where the experimental group's average t-unit length increased by almost two words in contrast to the the control group's average increase of only a fifth of a word.

2b Fluency

PHN

	Control Group	Exper. Group	t	p<0.05
n	(34)	(34)		
mean (of gain scores)	93.9	110.8	-0.77	0.45
unmatched t-test to compare PHN gain scores on average number of words per essay				

ISN

	Control Group	Exper. Group	t	p<0.05
n	(34)	(34)		
mean (of gain scores)	84.2	106	-0.98	0.33
unmatched t-test to compare ISN gain scores on average number of words per essay				

There were no significant differences on either PHN or ISN between the groups, but the experimental group increased their average number of words in both types more than the control group.

2c Accuracy

PHN - Total Error

	Control Group	Exper. Group	t	p<0.05
n	(34)	(34)		
mean (of gain scores)	-1.62	-2.04	0.48	0.63
unmatched t-test to compare PHN gain scores on average number of errors per 100 words				

ISN - Total Error

	Control Group	Exper. Group	t	p<0.05
n	(34)	(34)		
mean (of gain scores)	-1.64	-1.5	-0.17	0.86
unmatched t-test to compare ISN gain scores on average number of errors per 100 words				

There were no significant differences between the groups in the level of their error decrease in either PHN or ISN, but the experimental group reduced their error level more than the control group in PHN and less than the control group in ISN.